BAYESIAN INFERENCE ON MIXTURE MODELS AND THEIR APPLICATIONS

A Dissertation

by

ILSUNG CHANG

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

May 2005

Major Subject: Statistics

BAYESIAN INFERENCE ON MIXTURE MODELS AND THEIR APPLICATIONS

A Dissertation

by

ILSUNG CHANG

Submitted to Texas A&M University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Approved as to style and content by:

James A. Calvin (Co-Chair of Committee)	Bani K. Mallick (Co-Chair of Committee)
Ruzong Fan (Member)	
Thomas Schlumprecht (Member)	Simon Sheather (Head of Department)

May 2005

Major Subject: Statistics

ABSTRACT

Bayesian Inference on Mixture Models and Their Applications. (May 2005)

Ilsung Chang, B.S., Seoul National University;

M.S., Seoul National University

Co-Chairs of Advisory Committee: Dr. James A. Calvin
Dr. Bani K. Mallick

Mixture models are useful in describing a wide variety of random phenomena because of their flexibility in modeling. They have continued to receive increasing attention over the years from both a practical and theoretical point of view. In their applications, estimating the number of mixture components is often the main research objective or the first step toward it. Estimation of the number of mixture components heavily depends on the underlying distribution. As an extension of normal mixture models, we introduce a skew-normal mixture model and adapt the reversible jump Markov chain Monte Carlo algorithm to estimate the number of components with some applications to biological data.

The reversible jump algorithm is also applied to the Cox proportional hazard model with frailty. We consider a regression model for the variance components in the proportional hazards frailty model. We propose a Bayesian model averaging procedure with a reversible jump Markov chain Monte Carlo step which selects the model automatically. The resulting regression coefficient estimates ignore the model uncertainty from the frailty distribution. Finally, the proposed model and the estimation procedure are illustrated with simulated example and real data.

To my families and lovely E.K.

ACKNOWLEDGEMENTS

I would like to thank Dr. James A. Calvin, one of my co-advisors for his guidance and support throughout the work leading to this dissertation. My sincere gratitude and appreciation is given to Dr. Bani K. Mallick, my other co-advisor. His encouragement and support have made me better than I could be. I thank both of them for all they have taught me from the beginning of the doctoral courses at Texas A&M. Working with them has been nothing less than an honor.

I also wish to thank my committee members, Dr. Ruzong Fan and Dr. Thomas Schlumprecht, for their advice and help during my studies. I am also indebted to the Department of Statistics at Texas A&M for their support since I arrived in College Station. Special thanks are given Dr. Johan Lim, a friend of mine, for his suggestions.

I wish to express my deep gratitude to my parents, my parents-in-law, and sisters and brothers for their encouragement and their prayers. Of all people, I am most thankful to my wife, Eunkyung, who has been the greatest blessing in my life.

Thanks be to the Lord Jesus Christ for guiding me in the completion of this dissertation.

TABLE OF CONTENTS

Page

CHAPTER Page

LIST OF FIGURES

FIGURE Page

FIGURE Page

FIGURE Page

LIST OF TABLES

TABLE Page

CHAPTER I

INTRODUCTION

The mixture model has been extensively studied and challenged in various scientific fields such as genetics, biology, economics and other fields since a series of papers by Karl Pearson (1894; 1895). Inference on the number of mixture components is often the main research objective or could be the very first step for other inferential procedures. But it has been observed that the result depends on distributional assumptions (MacLean et al., 1976; Richardson and Green, 1997). For example, when the inherent distributional skewness is commingled with heterogeneity, the conventional normal mixture models, the most common distributional assumption, can overestimate the number of mixture components.

As an alternative to normal mixture model when the normality assumption seems violated, Stephens (1997) and Peel and McLachlan (2000) considered a mixture of t-distributions to address the heavy-tailed one. Taking transformations, e.g. log transformation, has also been applied to remove or reduce skewness in the data, for example, log transformation. The transformation may not be considered as merely a measure of how far it is from normality since it changes the original unit of the data, which implies a careful interpretation on the results. In this dissertation, we suggest changing the underlying distribution to the skew normal distribution (Azzalini, 1985), which includes the normal distribution as a special case. Loosening the normality

---

The format and style follow that of the *Journal of the American Statistical Association*.

assumption by allowing a skewness parameter in the normal distribution, the mixture of the skew normal ones is shown to be useful in inference on the mixture model. Also it is more realistic and helpful to analyze the data on the original scale, the scale by which the data are measured, in some cases such as in the prediction of a future observation.

The reversible jump algorithm by Green (1995) has been developed in cases where the dimension of the unknowns is itself unknown, which is the case here, the unknown number of components in the mixture model. Throughout the dissertation, we focus on the skew-normal mixture model as well as on how to apply the reversible jump algorithm to the proportional hazard model with frailty. The dissertation is organized as follows.

In Chapter II, we introduce the univariate skew-normal mixture model and its applications followed by multivariate skew-normal mixture in Chapter III. In Chpater IV, Bayesian model averaging is applied for the analysis on the heterogeneous frailty model along with the reversible jump algorithm followed by Chapter V that summarizes this study.

CHAPTER II

THE UNIVARIATE SKEW-NORMAL MIXTURE

2.1 Introduction

The mixture model has been extensively studied and challenged in various scientific fields such as genetics, biology, economics and other fields since a series of papers by Karl Pearson (1894; 1895). It has also had great effect on the various area of statistics, for example, discriminant analysis, clustering and latent class analysis (McLachlan and Peel, 2000). In these examples, inference on the number of mixture components is often the main research objective or could be the very first step for other inferential procedures.

It has been observed that inference on the number of components depends on distributional assumptions (MacLean et al., 1976; Richardson and Green, 1997). For example, when the inherent distributional skewness is commingled with heterogeneity, the conventional normal mixture models, the most common distributional assumption, can overestimate the number of mixture components. Pearson (1895) noted that an important problem is how we are to discriminate between a true curve of skew type and a compound curve, supposing we have no reason to suspect our data a priori of mixture. The question of how many components are needed to describe a phenomenon has been addressed in various techniques (Titterington et al., 1985; McLachlan and Peel, 2000), most of which are focused on the normal mixture model. The debate about the blood pressure distribution in 1960's is an example that shows the effect of the choice of the underlying distribution on the conclusion of the number

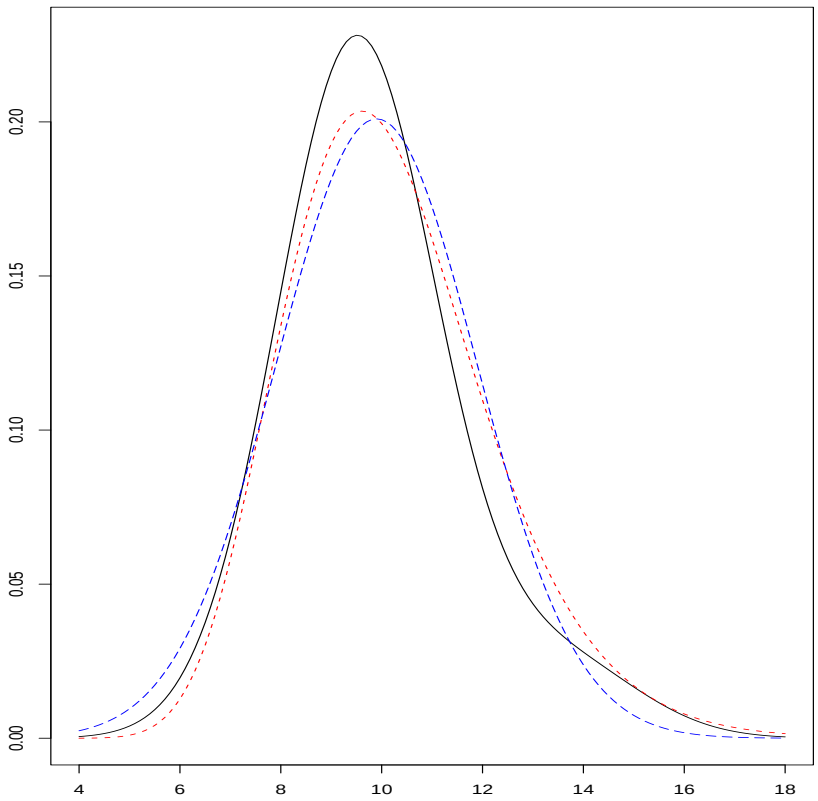

Figure 1: Plots of 3 different densities: mixture of two normals in solid line, log normal density in dashed line and normal density in dotted line.

of components needed to describe a phenomenon. See Schork et al. (1990) for the details. Another example of the difficulty in estimating the number of components in a mixture is illustrated in Figure 1 which is reproduced from Example 2.2.9 in Titterington et al. (1985) with one normal density added. A mixture of two normals is represented by the solid line, the log normal density by the dashed line and a normal density by the dotted line. All are so close, implying that it will very difficult to discriminate between them in practice.

When a heterogeneous population is a priori postulated and one is interested in

estimating the number of heterogeneous groups from which a random sample is taken, the distributional assumption governing the entire statistical procedure strongly affects the estimation of the number of components. There is a need to check this distributional assumption, but it is not easily achieved when there is skewness or a heavy-tailed distribution, as is seen in Figure 1. As an alternative to normal mixture model when the normality assumption seems violated, Stephens (1997) and Peel and McLachlan (2000) considered a mixture of t-distributions to address the heavy-tailed one. Taking transformations, e.g. log transformation, has also been applied to remove or reduce skewness in the data, for example, log transformation.

MacLean et al. (1976) proposed the procedure of the simultaneous estimation of a skewness parameter with mixture model parameters assuming a homogeneous variance. The transformation they chose is a scaled version of the power transform by Tukey (1957) and Box and Cox (1964). Schork and Schork (1988) proposed a method of a multivariate generalization of the technique of MacLean et al. (1976) allowing for unequal variance-covariance structures among the component distributions.

Gutierrez et al. (1995) successfully applied the Box-Cox transformation to estimate how many cells in tomato parent roots initiate a lateral root. The analysis on the original scale of the data suggests more than one group, a group which initiates a lateral root from two cells and another group which initiates from three cells in parent roots. But the inverse transformation suggests a different conclusion that one group is enough to explain the lateral root initiation from the parent root. When non-positive observations exist, variables should be translated via a location parameter to assume positive values before a power transformation is taken. MacLean et al. (1976) chose to estimate this location parameter along with the power or skewness parameters. Schork and Schork (1988) assume this location parameter is constant as established with discretion by the user before analysis. Permitting this location

parameter to vary in a numerical routine to find maximum likelihood estimates for parameters has been reported to increase the numerical instability of convergence while adding an unimportant parameter to be estimated. Liu et al. (2003) applied the Box-Cox transformation using a Markov chain Monte Carlo algorithm in the mixture model. They used principal component analysis for dimension reduction and estimated a power transform along with a translation parameter to make the data all positive. An alternative transformation is the exponential data transformation by Manly (1976) which does not need the data to be all positive. Richardson and Green (1997) applied it to a skewed data which suggests more than three components in the original scale but with Manly's transformation the posterior probability is maximized with two components.

In the chapter, we suggest changing the underlying distribution to the skew normal distribution (Azzalini, 1985), which includes the normal distribution as a special case. Loosening the normality assumption by allowing a skewness parameter in the normal distribution, the mixture of the skew normal ones is shown to be useful in inference on the mixture model. It is more realistic and helpful to analyze the data on the original scale, the scale by which the data are measured, in some cases such as in the prediction of a future observation.

For a criterion for the choice of the number of components, we adopt reversible jump algorithm of Green (1995) in this chapter. It generalizes the traditional Markov chain Monte Carlo algorithm to the case where the dimension of the unknown parameters in the model is also unknown. Richardson and Green (1997) applied the reversible jump MCMC method to the mixture of normal distributions with an unknown number of components. The algorithm produces the posterior probability of the number of components, upon which one draws a conclusion on how many components are needed to describe the data. After an introduction to the basic of mixture

model and skew normal distribution in Section 2.2 along with a short review of the applications of skew normal distribution in the literature, Section 2.3 describes the procedure of the inference in the mixture of skew normal ones following the basic introduction to the reversible jump algorithm by (Green, 1995). The application to the real data is explained in Section 2.4 followed by the conclusion and discussion in Section 2.5.

2.2 Models

Suppose that a random variable or vector, X, takes values in a sample space, $\mathcal{X}$, and that its distribution can be represented by a probability density function (or mass function in the case of discrete $\mathcal{X}$) of the form

$$p(x|\boldsymbol{\psi}) = \pi_1 f_1(x|\theta_1) + \pi_2 f_2(x|\theta_2) + \cdots + \pi_k f_k(x|\theta_k) \quad (x \in \mathcal{X}), \tag{2.1}$$

where $\boldsymbol{\psi}$ denotes all the parameters in the model, θ_j is component specific parameters, and

$$\sum_{j=1}^{k} \pi_j = 1, \quad \int_{\mathcal{X}} f_j(x|\theta_j)dx = 1, \quad \pi_j > 0, \quad f_j(\cdot) \geq 0, \ j = 1, ..., k.$$

Here we assume that X is a real-valued random variable. The parameters $\pi_1, ..., \pi_k$ are often called the mixing weights and $f_1(\cdot), ..., f_k(\cdot)$ the component densities of the mixture. We shall denote the collection of all distinct parameters occurring in the component densities by $\boldsymbol{\theta}$, and the complete collection of all distinct parameters occurring in the mixture model by $\boldsymbol{\psi}$ so that $\boldsymbol{\psi} = (\boldsymbol{\theta}, \boldsymbol{\pi}), \ \boldsymbol{\pi} = (\pi_1, ..., \pi_k)$ following the notation in Titterington et al. (1985). The observations $x_1, ..., x_n$ can be viewed as being incomplete. Suppose that the measurements are available from experimental units which are known to belong to one of a set of classes, but whose individual class-memberships are unavailable. Let the group label $\boldsymbol{z}_i$ be a vector of size k for the ith observation. If the ith observation x_i comes from the jth class, then

$z_{ij} = 1,\ z_{il} = 0,\ \ l \neq j$ with the probability of an observation x_i from the jth class equal to π_j. Given $\boldsymbol{z}_i$, the conditional distribution of x_i is, then,

$$x_i|\boldsymbol{z}_i \sim \prod_{j=1}^{k} f_j(x_i|\theta_j)^{z_{ij}} \quad \text{independently for } i = 1, ..., n,$$

thus,

$$f(x_i) = \sum_{j=1}^{k} p(x_i, z_{ij} = 1) = \sum_{j=1}^{k} p(z_{ij} = 1) f(x_i|z_{ij} = 1) = \sum_{j=1}^{k} \pi_j f_j(x_i|\theta_j).$$

Let $\boldsymbol{x}^n = (x_1, ..., x_n)^\top$ and $n \times k$ matrix $\boldsymbol{z}^n = (\boldsymbol{z}_1, ..., \boldsymbol{z}_n)^\top$, then, the complete likelihood function $f(\boldsymbol{x}^n, \boldsymbol{z}^n)$ is

$$L(\boldsymbol{x}^n, \boldsymbol{z}^n|\boldsymbol{\psi}) = \prod_{i=1}^{n} \prod_{j=1}^{k} \{\pi_j f_j(x_i|\theta_j)\}^{z_{ij}}. \tag{2.2}$$

While the observed likelihood obtained from the equation (2.1) is composed of the product of k sums, the complete likelihood of the equation (2.2) is only involved with the product of each component density and mixing weights. It is, therefore, straightforward to proceed the inference, for example, finding the maximum likelihood estimates.

In the case where the number of components is unknown, the notation may call for some conflicts as pointed out in Nobile (2004) since the meaning of mixture weights and mixture components is completely specified only when k is fixed: for instance, the expression "the weight of the second component" seems to have a different meaning when $k = 2$ than when, for example, $k = 5$. Keeping in mind that k is unknown, we follow the notation in the equation (2.1) for convenience.

In many cases the underlying density $f_j(\cdot|\theta_j)$ is assumed to be normal. In the case of certain biological, sociological and economic measurements there is a well-marked deviation from the normal shape, and it becomes important to determine the direction and amount of such deviation. The asymmetry may arise from the fact

that the units grouped together in the measured material are not really homogeneous. (Pearson, 1894). Including the normal family as a special case, we suggest to use the skew-normal distribution which is introduced in the following subsection.

2.2.1 Skew-normal distribution

The skew-normal distribution is first named by Azzalini (1985) but its appearance in the literature dates back to Roberts (1966). Let X be a skew normal distribution with skewness parameter λ, denoted by $SN(\lambda)$, then its probability density function, $\phi_\lambda(x)$, is

$$\phi_\lambda(x) = 2\phi(x)\Phi(\lambda x), \tag{2.3}$$

where $\phi(x)$ and $\Phi(x)$ denote the density and the distribution function of the standard normal random variable, respectively. The skew-normal distribution is a special case of generalized skew-elliptical (GSE) distributions defined by Genton and Loperfido (2005). The moment generating function $M_\lambda(t)$ is

$$M_\lambda(t) = 2\exp(t^2/2)\Phi(\delta t), \qquad \delta = \frac{\lambda}{\sqrt{1+\lambda^2}}, \tag{2.4}$$

from which we know $E(X) = b\delta, E(X^2) = 1, E(X^3) = 3b\delta - b\delta^3$ with $b = \sqrt{2/\pi}$. The index of skewness denoted by γ^3 is

$$\gamma^3(X) = E\left(\frac{X - EX}{\sqrt{var(X)}}\right)^3 = \sqrt{2}(4-\pi)\left(\frac{\lambda}{\sqrt{(\pi + (\pi-2)\lambda^2)}}\right)^3.$$

Figure 2 illustrates some of skew-normal densities in (2.3) over different values of the skewness parameter λ. When $\lambda = 0$, it is equivalent to the standard normal distribution and it becomes more skewed to the right as λ goest to ∞ or skewed to the left when λ goes to $-\infty$. With $\lambda = \infty$, it is half normal distribution on the positive support or vice versa with $-\infty$ of λ.

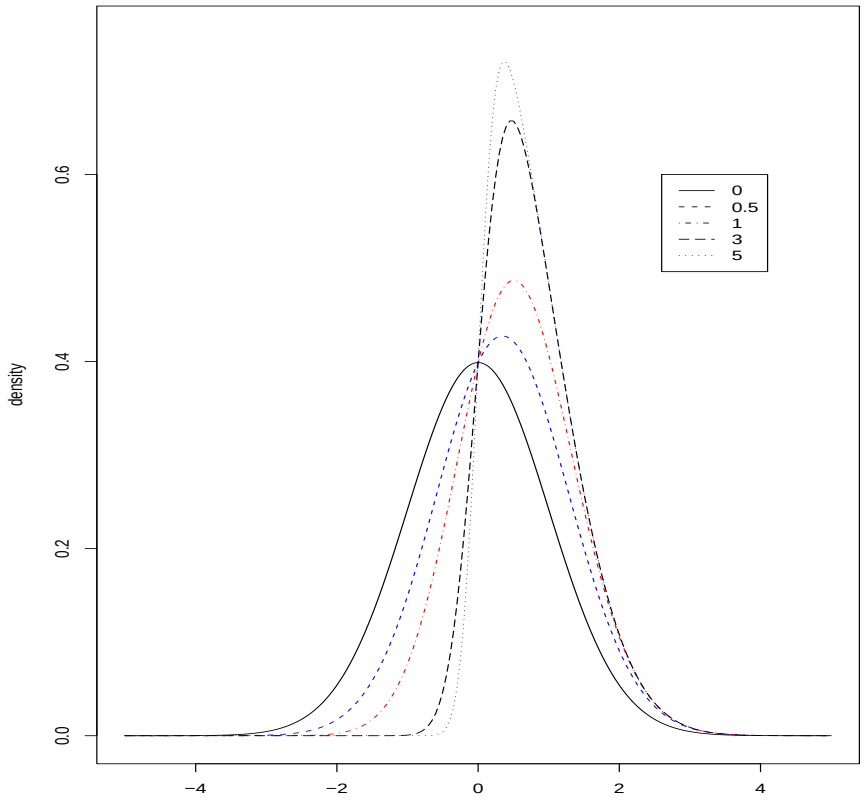

Figure 2: Plots of skew-normal densities with different value of λ: $\lambda = 0$ to $\lambda = 5$ from lowest to highest.

The distribution can be generalized by the inclusion of location and scale parameters. If $X \sim SN(\lambda)$, then $Y = \xi + \omega X$ is a skew-normal random variable denoted by $SN(\xi, \omega, \lambda)$ with

$$\mu = E(Y) = \xi + \omega b\delta, \quad \sigma^2 = E(Y^2) = \xi^2 + 2\xi\omega b\delta,$$

$$E(Y^3) = \xi^3 + 3\xi^2\omega b\delta + 3\xi\omega^2 + 3\omega^3 b\delta - \omega^3 b\delta^3.$$

and its density has the form of

$$f(y) = 2(2\pi\omega^2)^{-1/2} \exp\left[-\frac{1}{2}\frac{(y-\xi)^2}{\omega^2}\right] \Phi\left(\frac{\lambda(y-\xi)}{\omega}\right). \tag{2.5}$$

Note that the coefficient of skewness for Y is the same as that for X. Such a parameterization is called the direct parameterization as in Azzalini and Capitanio (1999). Our approach for the inference is Bayesian. For the inference of the parameters by maximum likelihood or EM algorithm, refer to Azzalini and Capitanio (1999), where the centered parameterization is also mentioned.

It considers (μ, σ^2) instead of (ξ, ω) and will be used in the reversible jump algorithm later in the moment matching conditions that will be illustrated in section 3.3.

2.2.2 SN in the literature

Roberts (1966) obtained the form of skew-normal density in the context of the bivariate normal distribution. Assume that (X, Y) is a standard bivariate normal distribution with correlation ρ. The random variables (X_i, Y_i), $i = 1, ..., n$ are independent observations on (X, Y). Let $Z_i = \min(X_i, Y_i)$. Suppose that one is not able to observe X's or Y's and observes only the Z's. Then the density of Z is

$$f(z) = 2\phi(z)\Phi(-z\sqrt{(1-\rho)/(1+\rho)}).$$

In general, a function that is the product of a density and a distribution function can be a density under a certain conditions. Let f be a density function symmetric at 0, and G an absolutely continuous distribution function such that $G^{'}$ is symmetric at 0. Then

$$2f(x)G(\lambda x) \quad (y \in R)$$

is a density function for any real λ . In the skew-normal distribution ϕ is chosen for f and Φ for G. For the case of Roberts (1966), $\lambda = -\sqrt{(1-\rho)/(1+\rho)}$.

Arnold et al. (1993) developed the skew normal distribution in a different set up as a marginal distribution of a truncated bivariate normal distribution. This structure will be used for Bayesian parameter estimation using MCMC explained in section 3.3 in detail.

Suppose f is a bivariate normal density with zero mean vector and unit variances with correlation ρ and that (X,Y) has joint density $f_{X,Y}(x,y) \propto f(x,y)I(a < y < b)$, where a and b are real constants that are the lower and upper truncation points for Y. The normalizing constant is the inverse of $\Phi(b) - \Phi(a)$. Then the marginal distribution of x is, by direct integration,

$$f_X(x) = \phi(x)\frac{\Phi(\frac{b-\rho x}{\sqrt{1-\rho^2}}) - \Phi(\frac{a-\rho x}{\sqrt{1-\rho^2}})}{\Phi(b) - \Phi(a)} \quad (x \in R)$$

As expected, when $\rho = 0$ it density becomes normal. For $a = 0$ and $b = \infty$, it becomes the skew normal distribution with the skewness parameter $\lambda = \rho/\sqrt{1-\rho^2}$.

In a Bayesian framework, the skew-normal distribution was obtained by O'Hagan and Leonard (1976) as a prior distribution under hierarchical structure with constraints on the hyperparameter. Chen et al. (1999) introduce a new class of skewed link models for binary response data. They show that the underlying latent variable has a marginal distribution of a standard skew-normal distribution under a certain distributional assumption on the model. To handle the skewed spatial data, Kim and

Mallick (2004) present a model based on the skew normal distribution. Azzalini and Dalla Valle (1996) extend the skew normal distribution to the multivariate situation. Azzalini and Capitanio (1999) presents an successful example of application of the multivariate skew normal distribution to the real data. When the standard multiple regression is supplemented by a 'selection equation' of the quantity that is not observed, the simplest case of the conditional distribution of the observation given the missing information is the skew-normal distribution as Copas and Li (1997) mentioned. Carroll et al. (2004) use the skew-normal distribution for a simulation study to test robustness of their model noticing that the density is reasonably skewed for any value of $\lambda \geq 5$.

The reader shall be referred to Genton (2004) for more examples and history of the skew-normal distribution.

The following calculation shows a property of skew-normal distribution which is observed in the above literature reviews and is useful in both generating the samples from the skew-normal distribution and estimating the parameters of the skew-normal distribution. Suppose $(X_0, X)^\top$ follows a left-truncated bivariate normal distribution denoted by $N_2(\mu, \Sigma)I(X_0 > \mu_0)$ with $\mu = (\mu_0, \xi)^\top$ and $V(X_0) = \sigma_0^2$, $V(X) = \omega^2$ and $Corr(X_0, X) = \rho$. Then, the marginal distribution of X is $\mathrm{SN}(\xi, \omega, \lambda)$ where $\lambda = \frac{\rho}{\sqrt{(1-\rho^2)}}$ and $\rho = \frac{\lambda}{\sqrt{1+\lambda^2}}$. Note that the density function of Y does not depend on μ_0 nor σ_0 so that we may assume $\mu_0 = 0$ and $\sigma_0 = 1$.

2.2.3 Estimation procedure

Let $x_1, ..., x_n$ be the observations from a skew-normal distribution. From the last lemma with $\mu_0 = 0$ and $\sigma_0 = 1$, we can generate a set of random variables $u_1, ..., u_n$ from a truncated normal distribution to estimate the parameters via EM type algo-

rithm (Dempster et al., 1977). The complete likelihood becomes

$$f(\theta; x^n, u^n) = \prod_{i=1}^{n} I_{(u_i>0)} 2(2\pi\omega\sqrt{1-\rho^2})^{-1} e^{-\frac{1}{2(1-\rho^2)}(u_i^2 - 2\rho u_i(x_i-\xi)/\omega + (x_i-\xi)^2/\omega^2)}.$$

We take a hierarchical structure of the priors following Richardson and Green (1997), that is,

$$\xi \sim N(a_1, 1/\kappa^{-1}), \quad \tau = \omega^{-2} \sim G(a_2, b_2), \quad b_2 \sim G(g, h), \quad \lambda \sim N(0, b_3^2),$$

thus, the posterior density is proportional to

$$\exp\left\{-\frac{1+\lambda^2}{2}\sum_i (u_i^2 - 2\frac{\lambda}{\sqrt{1+\lambda^2}} u_i(x_i-\xi)\tau^{0.5} + \tau(x_i-\xi)^2)\right\} \times$$
$$I(u_{(1)} > 0)\tau^{\frac{n}{2}+a_2-1} e^{-b_2\tau} b_2^{a_2+g-1} e^{-hb_2} e^{-\lambda^2/(2b_3^2)} (1+\lambda^2)^{\frac{n}{2}},$$

based on which we estimate the parameters as the following **Algorithm 1**.

Algorithm 1.

Given $\theta = (\xi, \tau, \lambda)^\top$ with $\omega = 1/\sqrt{\tau}, \quad \rho = \lambda/\sqrt{1+\lambda^2}$,

1. Generate $U_i \sim N(\rho(x_i - \xi)/\omega, (1-\rho^2)) \cdot I(U_i > 0), i = 1, ..., n$

2. Update ξ from $N(\mu_\xi, \sigma_\xi^2)$, where

$$\sigma_\xi^2 = (\kappa + n\tau(1+\lambda^2))^{-1}$$

$$\mu_\xi = \sigma_\xi^2 \times \left(a_1\kappa + \tau(1+\lambda^2)\sum_i x_i - \lambda\sqrt{\tau(1+\lambda^2)}\sum_i u_i\right)$$

3. Generate $(\tau^{(c)}, \lambda^{(c)})$ from a 2-dimensional proposal distribution and update via the usual Metropolis-Hastings algorithm.

4. Update b_2 from $G(a_2 + g, \tau + h)$.

2.3 Model Determination

In many cases, the number of components itself is in interest to be estimated. There are several methods to test the number of components, one of which is to use BIC suggested by Schwarz (1978) and implemented in the model-based clustering by Fraley and Raftery (2002). The number of component which yields the maximum BIC would be chosen for the estimate of unknown number of components. When it is possible to obtain the marginal likelihood conditional on the fixed number of components, Bayes Factor can be used for the model choice criterion. The following subsection illustrates how to calculate the marginal likelihood in both normal mixture and skew normal mixture. The reversible jump algorithm developed by Green (1995) and followed by Richardson and Green (1997) for the application to the finite mixture model is such an algorithm to estimate simultaneously the parameters of the component densities and the unknown number of components. Section 3.3 displays a short introduction to the reversible jump followed by the detail of the implementation into the mixture based on Richardson and Green (1997).

2.3.1 Bayes factor

Let M_1 and M_2 denote two different models in interest, and take a look at the posterior odds ratio

$$\frac{p(M_1|y)}{p(M_2|y)} = \frac{p(y|M_1)}{p(y|M_2)} \times \frac{p(M_1)}{p(M_2)},$$

where y denotes the set of observations and $p(M_i)$ is the prior probability of $M_i,\ i = 1, 2$ being the true model. To determine this ratio we need to multiply the prior odds ratio by what is known as the marginal likelihood ratio (also known as the integrated likelihood ratio of prior predictive) (Denison et al., 2002). The marginal likelihood of model M_i gives a measure of the probability of observing the data given that M_i

is true. To account for the uncertainty in the unknowns associated with each model we determine the marginal likelihood by integrating out the model parameters. The Bayes factor is defined for the comparison of two competing models and, if we wish to consider the relative merits of M_i over M_j, is given by

$$BF_{ij} \equiv (M_i, M_j) = \frac{p(M_i|D)}{p(M_j|D)} / \frac{p(M_i)}{p(M_j)},$$

the posterior to prior odds ratio. In the case where the prior probabilities of each model have been taken to be equal, we find that the Bayes factor is exactly the same as the posterior odds ratio. In this case, choosing the model with the highest posterior probability is equivalent to picking the model whose Bayes factor with respect to any other model is greater than one. Kass and Raftery (1995) suggest that, if the Bayes factor for M_i over M_j is between 1 and 3, then there is little perceived difference between the models, between 3 and 20 there is positive evidence in favor of M_i, 20 to 150 strong evidence and, if the Bayes factor is over 150, there is very strong evidence in favor of M_i.

Under the equal prior probabilities, to obtain Bayes factor reduces to calculate the marginal likelihood. Chib (1995) developed how to calculate the marginal likelihood from the Gibbs sampler outputs, and the normal mixture is one of the examples that are described.

Let $\boldsymbol{w}$ be the mixing weights and $\boldsymbol{\theta}$ be all the parameters in the component densities. Note that $p(y|M_1)$ can be written as

$$p(y|M_1) = \frac{f(y|\boldsymbol{w}, \boldsymbol{\theta}, M_1)\pi(\boldsymbol{w}, \boldsymbol{\theta}|M_1)}{p(\boldsymbol{w}, \boldsymbol{\theta}|y, M_1)},$$

where the numerator is just the product of the sampling density (the likelihood) and the prior, and the denominator is the posterior density of $\boldsymbol{w}$ and $\boldsymbol{\theta}$. This simple identity, which holds for any $\boldsymbol{w}$ and $\boldsymbol{\theta}$, is referred to as the basic marginal likelihood

identity (BMI). It is necessary to estimate the value of the denominator unless we know exactly the posterior density of the $\boldsymbol{\theta}$ and $\boldsymbol{w}$. For given $\boldsymbol{w}$ and $\boldsymbol{\theta}$ (say $\boldsymbol{w}^*$ and $\boldsymbol{\theta}^*$), all it requires is the calculation of the log likelihood function, the prior and an estimate of posterior ordinate $p(\boldsymbol{w}^*, \boldsymbol{\theta}^*|y, M_1)$ and it will be done with the collection of samples from the Gibbs algorithm.

The samples from the Gibbs algorithm can be used here to estimate the value of posterior density evaluated at $\boldsymbol{w}^*$ and θ^*. Although this procedure leads to an increase in the number of iterations, it is important to stress that it does not require new programming and thus is straightforward to implement. The estimate of the marginal density becomes, in log scale,

$$\log \hat{p}(y|M_1) = \log f(y|\boldsymbol{w}^*, \boldsymbol{\theta}^*, M_1) + \log \pi(\boldsymbol{w}^*, \boldsymbol{\theta}^*|M_1) - \log \hat{p}(\boldsymbol{w}^*, \boldsymbol{\theta}^*|y, M_1).$$

Chib and Jeliazkov (2001) showed how to calculate the marginal likelihood from the Metropolis-Hastings output, which is the case of the skew normal distribution because we use the Metropolis-Hastings algorithm to update the scale parameter and the skewness parameter.

2.3.2 Reversible jump MCMC

The reversible jump algorithm developed by Green (1995) solves dimension changing problem as change point analysis or mixture analysis. For details, refer to Chapter VI. in Green et al. (2003), where it is extended to the name of trans-dimensional MCMC to incorporate the situation where the 'unknowns' in the Bayesian set-up does not have fixed dimension.

Suppose that a move type m is proposed, from x, the present state vector with a lower dimension, to a point $x^{'}$, the destination in a higher dimensional space. This will very often be implemented by drawing a vector of continuous random variables u,

independent of x, and setting $x^{'}$ by using an invertible deterministic function $x^{'}(x,u)$ so that the dimension of (x,u) is equal to that of $x^{'}$. The reverse of the move (from $x^{'}$ to x) can be accomplished by using the inverse transformation, so that the proposal is deterministic. Then the acceptance probability of the move type m is

$$\min\left\{1, \frac{p(x^{'}|y)r_m(x^{'})}{p(x|y)r_m(x)q(u)}\left|\frac{\partial x^{'}}{\partial(x,u)}\right|\right\},$$

where $r_m(x)$ is the probability of choosing move type m when in state x, and $q(u)$ is the density function of u. Note that the final term in the ratio above is a Jacobian arising from the change of variable from (x,u) to $x^{'}$.

Figure 3 shows how the dimension varying move in the mixture model is obtained.

The following section describes the implementation of the reversible algorithm based on Richardson and Green (1997).

2.3.3 Moment matching condition

One of the advantages of the method is to model the number of components and the mixture component parameters jointly and base inference about these quantities on their posterior probabilities. This is in contrast with most previous Bayesian treatments of mixture estimation, which consider models for different numbers of components separately and use significance tests or other non-Bayesian criteria to infer the number of components. With the priors $p(k,\boldsymbol{\psi}) = p(k,\boldsymbol{w},\boldsymbol{\xi},\boldsymbol{\omega},\boldsymbol{\lambda})$, the following algorithm is used to implement the parameters for fixed k:

Algorithm 2.

Given $\boldsymbol{\psi}^{(t)}$

1. Update the group label z by

$$p(z_i = j|\cdots) \propto w_j SN(x_i \mid \theta_j).$$

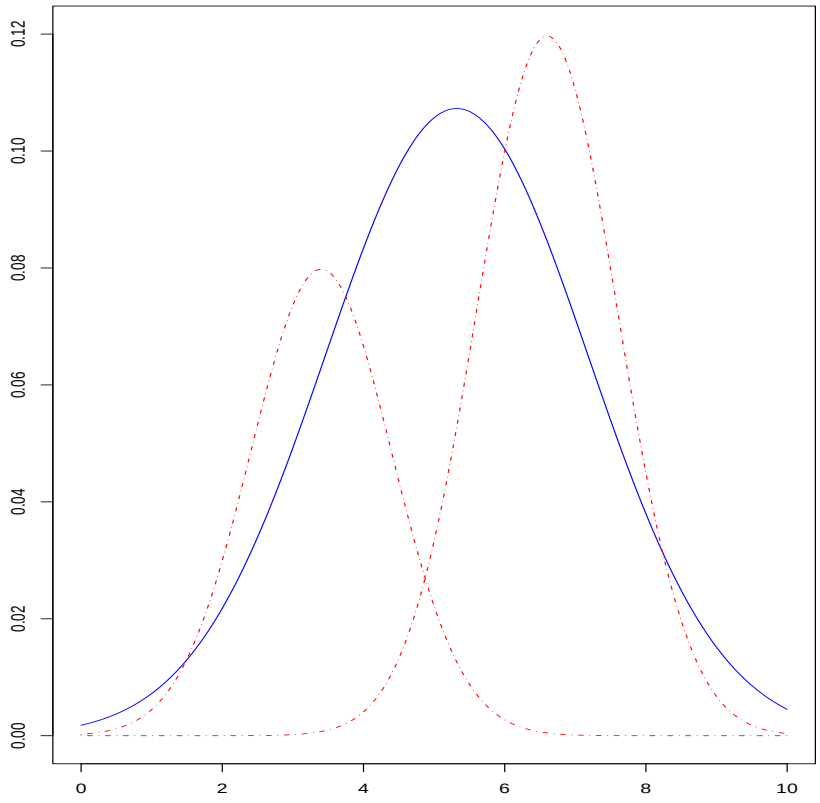

Figure 3: Sample plots of combining two groups to one or splitting one group into two in a dimesion changing move.

2. Allocate the observations to the appropriate group j, $j = 1, ..., k$ based on the group label z and calculate n_j, the number of elements in N_j.

3. Update the weights w by

$$w \sim Dirichlet(\alpha_1 + n_1, ..., \alpha_k + n_k),$$

where $Dirichlet(\alpha_1, ..., \alpha_k)$ is prior distribution for w.

4. For each $j = 1, ..., k$, apply **Algorithm 1** to update the parameters $\xi_j,\ \omega_j,\ \lambda_j$.

Before explaining the reversible jump algorithm, we need to define a set of parameters based on the moments of the skew-normal distribution. For up to 3 rd moments, we define $\mu,\ \sigma^2$ and γ^3 by

$$\begin{aligned} \mu &= E(X) &&= \xi + \omega b\delta, \\ \mu^2 + \sigma^2 &= E(X^2) &&= \xi^2 + 2b\xi\omega\delta + \omega^2, \\ \gamma^3 &= E\left(\frac{X - E(X)}{\sqrt{Var(X)}}\right)^3 &&= b(2b^2 - 1)\delta^3/(1 - b^2\delta^2)^{3/2}, \end{aligned}$$

where $b = \sqrt{2/\pi},\ \delta = \lambda/\sqrt{1 + \lambda^2}$ and note that ξ is the location parameter, ω the scale parameter, and λ the shape parameter in the directed parameterization. Here the following relation is observed:

$$\begin{aligned} \mu = &\quad \xi + \omega b\delta, &&\xi = \mu - \omega b\delta \\ \sigma^2 = &\quad \omega^2(1 - b^2\delta^2), &&\omega = \frac{\sigma}{\sqrt{1 - b^2\delta^2}} \\ \gamma = &\quad cb\delta/\sqrt{1 - b^2\delta^2}, &&\lambda = \frac{\delta}{\sqrt{1 - \delta^2}}, \end{aligned} \tag{2.6}$$

where $c^3 = (4 - \pi)/2$ and $\delta = \lambda/\sqrt{1 + \lambda^2} = 1/b \cdot \gamma/\sqrt{c^2 + \gamma^2}$. Note that as δ goes to 1, γ^3 goes to $\sqrt{2}(4 - \pi)/(\pi - 2)^{1.5} \sim .9952717 \equiv R_3$.

To update the number of component k, we combine any two adjacent components into one or split a component chosen randomly into two, and then reallocate the observations x_i, rescale the weight parameters, and assign the parameters $\theta = (\xi, \omega, \lambda)$

using the moment matching condition. First, we make a random choice between attempting to split or combine, with probabilities b_k and $d_k = 1 - b_k$ respectively, depending on k. If the Poisson prior is considered for the parameter k, $d_1 = 0 = 1 - b_1$ and $b_k = d_k = 0.5$ for $k = 2, 3, ...$ If a discrete uniform distribution or a truncated Poisson distribution is considered, then we set $b_{k_{max}} = 0 = 1 - d_{k_{max}}$, where k_{max} is the maximum value allowed for k, additionally. The combine proposal will be done as follows: choose a pair of components (r, s) at random, that are adjacent in terms of the current value of their location parameters, i.e.

$$\xi_r < \xi_s, \quad \text{with no other } \xi_j \text{ in the interval } [\xi_r, \xi_s]. \tag{2.7}$$

These components are merged, reducing k by 1. In doing so, forming a new component here labeled $*$, we have to reallocate all those observations x_i with $z_i = r$ or $z_i = s$ and create values for (w_*, θ_*). The allocation is simply done by setting such $z_i = *$, where as the other parameters are assigned, through the above transformation, by the expedient of matching up to 3rd moments of the new component to those of a combination of the two that it replaces:

$$w_* = w_r + w_s \tag{2.8}$$

$$w_*\mu_* = w_r\mu_r + w_s\mu_s \tag{2.9}$$

$$w_*(\mu_*^2 + \sigma_*^2) = w_r(\mu_r^2 + \sigma_r^2) + w_s(\mu_s^2 + \sigma_s^2) \tag{2.10}$$

$$w_*(\mu_*^3 + 3\mu_*\sigma_*^2 + \sigma_*^3\gamma_*^3) = w_r(\mu_r^3 + 3\mu_r\sigma_r^2 + \sigma_r^3\gamma_r^3) + w_s(\mu_s^3 + 3\mu_s\sigma_s^2 + \sigma_s^3\gamma_s^3). \tag{2.11}$$

The reversible split proposal begins by choosing a component $*$ randomly and needs a four-dimensional random vector u for the parameter update. Let $u_1 \sim be(2, 2)$, where $be(a, b)$ has the density proportional to $u^{a-1}(1-u)^{b-1}$. From the equation (2.8) set $w_r = u_1 * w_*$ then $w_s = (1 - u_1)w_*$. From (2.9) we obtained $\mu_s = (w_*\mu_* - w_r\mu_r)/w_s$ and

by substituting it into (2.10), after some algebra, we easily noticed that the following equation holds:

$$\frac{w_s}{w_r}\sigma_*^2 = (\mu_r - \mu_*)^2 + \frac{w_s}{w_* w_r}(w_r\sigma_r^2 + w_s\sigma_s^2),$$

where the right hand side is the sum of the positive quantities and the left hand side is positive, thus, by letting $u_2 \sim be(2,2)$, we can have

$$\{\mu_r, \mu_s\} = \{\mu_* - u_2\sigma_*\sqrt{w_s/w_r}, \mu_* + u_2\sigma_*\sqrt{w_r/w_s}\}.$$

To bring both μ_r and μ_s into (2.10) again gives $(1-u_2^2)w_*\sigma_*^2 = w_r\sigma_r^2 + w_s\sigma_s^2$ so that for $u_3 \sim be(1,1)$ we have

$$\sigma_r^2 = u_3(1-u_2^2)\frac{w_*}{w_r}\sigma_*^2$$
$$\sigma_s^2 = (1-u_3)(1-u_2^2)\frac{w_*}{w_s}\sigma_*^2.$$

For the parameter γ^3s, since we have the restriction on the range of γ^3, instead of using the equation (2.11) and other parameters for the moment matching condition, we create a simple equation to generate the skewness parameter λ with u_4 from $N(0,\epsilon^2)$

$$\lambda_r = \lambda_* - u_4, \quad \lambda_s = \lambda_* + u_4.$$

After calculating $(w_r,\ w_s,\ \mu_r,\ \mu_s,\ \sigma_r^2,\ \sigma_s^2,\ \lambda_r^3,\ \lambda_s^3)$, we convert μ and σ into the directed parameterization by the relation (2.6), i.e.

$$(\mu_r,\ \mu_s,\ \sigma_r^2,\ \sigma_s^2) \to (\xi_r,\ \xi_s,\ \omega_r^2,\ \omega_s^2).$$

At this point, we check whether the adjacency condition (2.7) is satisfied. If not, the move from k to $k+1$ is rejected, as the pair could not then be reversible. If the test is passed, it remains only to propose the reallocation of those x_i with $z_i = *$ between r and s. This is done analogously to the standard Gibbs allocation move. See **Algorithm 2**.

The acceptance probability for the split move is $\min(1, A)$, where

$$\begin{aligned} A &= \frac{w_r^{L_r}\prod_{i\in N_r} SN(x_i;\xi_r,\omega_r,\lambda_r) w_s^{L_s}\prod_{i\in N_s} SN(x_i;\xi_s,\omega_s,\lambda_s)}{w_*^{L_*}\prod_{i\in N_*} SN(x_i;\xi_*,\omega_*,\lambda_*)} \cdot \frac{p(k+1)}{p(k)} \\ &\cdot \frac{\frac{\Gamma((k+1)\alpha)}{\Gamma(\alpha)^{k+1}} w_r^{\alpha-1} w_s^{\alpha-1}}{\frac{\Gamma(k\alpha)}{\Gamma(\alpha)^k} w_*^{\alpha-1}} \frac{p(\xi_r)p(\xi_s)}{p(\xi_*)} \frac{(k+1)!}{k!} \frac{p(\omega_r)p(\omega_s)}{p(\omega_*)} \frac{p(\lambda_r)p(\lambda_s)}{p(\lambda_*)} \\ &\cdot \frac{d_{k+1}}{b_k P_{alloc}} \frac{1}{p(u_1)p(u_2)p(u_3)p(u_4)} \text{ Jacobian,} \end{aligned}$$

where

$$N_r = \{i : z_i = r\}, \quad L_r = |N_r|, \quad N_s = \{i : z_i = s\}, \quad L_s = |N_s|,$$

$$N_* = \{i : z_i = *\} = N_r \cup N_s, \quad L_* = L_r + L_s$$

and, the Jacobian arises at the transformation

$$T_* = (w_*, \xi_*, \omega_*^2, \lambda_*, u_1, u_2, u_3, u_4) \ \rightarrow T_{rs} = (w_r, w_s, \xi_r, \xi_s, \omega_r^2, \omega_s^2, \lambda_r, \lambda_s).$$

Let $V_* = (w_*, \mu_*, \sigma_*^2, \lambda_*, u_1, u_2, u_3, u_4)$ and V_{rs}=(w_r,w_s,μ_r,μ_s,σ_r^2,σ_s^2,λ_r,λ_s). Then,

$$\begin{aligned} \text{Jacobian} &= \left|\frac{dT_{rs}}{dT_*}\right| = \left|\frac{dT_{rs}}{dV_{rs}}\right| \left|\frac{dV_{rs}}{dV_*}\right| \left|\frac{dV_*}{dT_*}\right| \\ &= \frac{1-b^2\delta_r^2}{\sigma_r^2}\frac{1-b^2\delta_s^2}{\sigma_s^2} \times \frac{\sigma_*^2}{1-b^2\delta_*^2} \times \frac{w_*|\mu_r-\mu_s|\sigma_r^2\sigma_s^2}{u_2(1-u_2^2)u_3(1-u_3)\sigma_*^2} 2 \\ &= \frac{(1-b^2\delta_r^2)(1-b^2\delta_s^2)}{1-b^2\delta_*^2}\frac{\sigma_*^2}{\sigma_r^2\sigma_s^2}\frac{2w_*|\mu_r-\mu_s|}{u_2(1-u_2^2)u_3(1-u_3)\sigma_*^2} \end{aligned}$$

and

$$\begin{aligned} P_{alloc} &= \prod_{i\in N_r} \frac{w_r SN(x_i;\xi_r,\omega_r,\lambda_r)}{w_r SN(x_i;\xi_r,\omega_r,\lambda_r) + w_s SN(x_i;\xi_s,\omega_s,\lambda_s)} \\ &\quad \cdot \prod_{i\in N_s} \frac{w_s SN(x_i;\xi_s,\omega_s,\lambda_s)}{w_r SN(x_i;\xi_r,\omega_r,\lambda_r) + w_s SN(x_i;\xi_s,\omega_s,\lambda_s)} \\ &= \frac{\prod_{i\in N_r} w_r SN(x_i;\xi_r,\omega_r,\lambda_r)\prod_{i\in N_s} w_s SN(x_i;\xi_s,\omega_s,\lambda_s)}{\prod_{i\in N_*}(w_r SN(x_i;\xi_r,\omega_r,\lambda_r) + w_s SN(x_i;\xi_s,\omega_s,\lambda_s))} \end{aligned}$$

For the combine move, the acceptance probability is $\min(1, A^{-1})$, using the same expression for A but with some obvious differences in the substitutions.

2.3.4 Prior specification

In a mixture context, being fully non-informative and obtaining proper posterior distributions are not possible. Since there is always the possibility that no observations are allocated to one or more components, and so the data are uninformative about them, standard choices of independent improper non-informative prior distributions for the component parameters cannot be used (Diebolt and Robert, 1994). We follow the prior suggested by Richardson and Green (1997) for the location and scale parameter. The prior for ξ_j is taken to be $N(a_1, \kappa^{-1})$, which seems natural to be rather flat over an interval of variation of the data, either postulated a priori or corresponding to the observed range. This can be achieved in a simple way by letting a_1 equal the midpoint of this interval, and setting κ equal to a small multiple of $1/R^2$, where R is the length of the interval. In contrast with the case of the location parameter, it seems restrictive to suppose that knowledge of the range of the data implies much about the size of the τ_j^{-2}. They introduce an additional hierarchical level by allowing b_3 to follow a gamma distribution with parameters g and h. Following Richardson and Green (1997) we take $a_2 > 1 > g$ to express the belief that the τ_j^{-2} is similar, without being informative about their absolute size. The scale parameter h will be a small multiple of $1/R^2$. The standard deviation, b_3, for the skewness parameter indicates the belief of how much the data is close to the normal. We assume it is not far from zero so that we take b_3 for small positive number, for example, 5 for the analysis of tomato data set instead of bigger number like 100 or 10000. Under certain prior structure, the posterior of the number of components is highly correlated with the specified values of the hyperparameters as discussed by Jennison (1997). The above hierarchical prior structure is shown to make the posterior probability of k insensitive. ¿From small simulation experience we notice that the choice of b_3 could

affect the posterior probability of k but our choice of b_3 seems to be reasonable.

2.3.5 Predictive distribution and classification probability

In Bayesian procedure we can obtain the predictive density estimate and the classification probability using MCMC samples. The predictive distribution summarizes the information concerning the likely value of a new observation, given the likelihood, the prior, and the data we have observed so far.

With the simulated samples we can obtain the following quantities: The density of a future observation given $\boldsymbol{x}^n$ conditional on fixed k is

$$p(x_{n+1}|\boldsymbol{x}^n, k) = \int p(x_{n+1}|\boldsymbol{x}^n, \boldsymbol{\psi}, k)p(\boldsymbol{\psi}|\boldsymbol{x}^n, k)d\boldsymbol{\psi} \approx \frac{1}{T}\sum_{t=1}^{T} p(x_{n+1}|\boldsymbol{\psi}^{(t)}, k), \quad (2.12)$$

the classification probability of the observed data is

$$\begin{aligned} Pr(z_{ij} = 1|\boldsymbol{x}^n, k) &= \int \frac{w_j SN(x_i|\theta_j)}{\sum_{l=1}^{k} w_j SN(x_i|\theta_l)} p(\boldsymbol{\psi}|\boldsymbol{x}^n, k)d\boldsymbol{\psi} \\ &\approx \frac{1}{T}\sum_{t=1}^{T} \frac{w_j^{(t)} SN(x_i|\theta_j^{(t)})}{\sum_{l=1}^{k} w_l^{(t)} SN(x_i|\theta_l^{(t)})}, \end{aligned} \quad (2.13)$$

and the classification probability of a future observation x_{n+1} is

$$Pr(z_{n+1,j} = 1|x_{n+1}, k) \approx \frac{1}{T}\sum_{t=1}^{T} \frac{w_j^{(t)} SN(x_{n+1}|\theta_j^{(t)})}{\sum_{l=1}^{k} w_l^{(t)} SN(x_{n+1}|\theta_l^{(t)})}. \quad (2.14)$$

2.4 Application

In this section, we apply the mixture of skew normal distribution under the number of component being unknown to the simulated data set as well as two real data sets. One is called tomato data set analyzed in Gutierrez et al. (1995) and the other one is enzyme data set analyzed in Bechtel et al. (1993). The explanation to the data set is described later in detail.

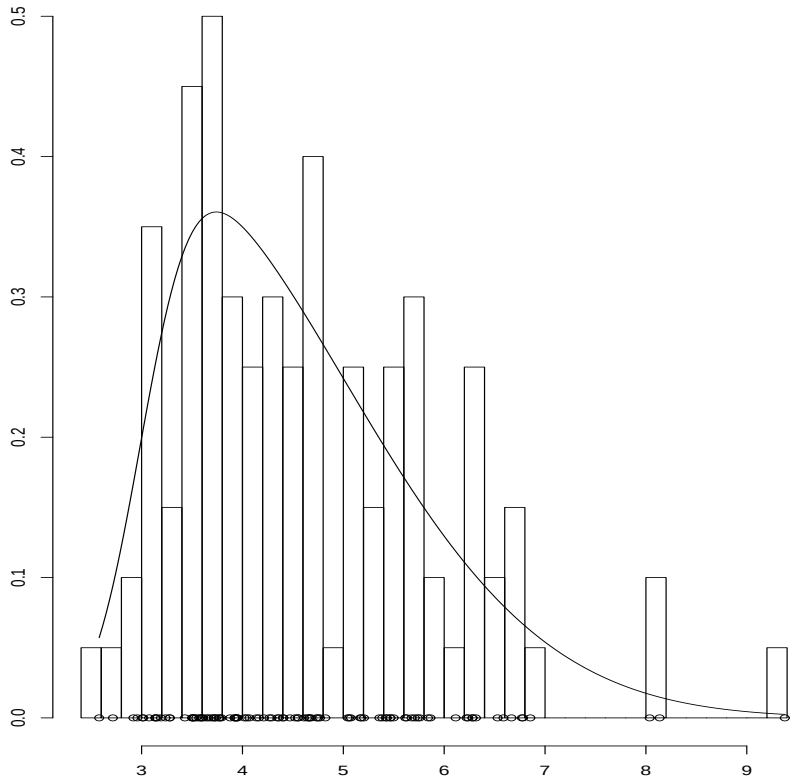

Figure 4: Histogram of 100 samples from SN(3,2,5). The circles on the horizontal axis denote the observations.

2.4.1 Simulation results

We generate $n = 100$ random numbers $x_1, ..., x_n$ from $SN(3, 2, 5)$. Figure 4 shows its histogram with the points on the x-axis. First, the appropriate power was sought conditional on k using its likelihood

$$L(\lambda, \boldsymbol{x}^n) = \prod_{i=1}^{n} x_i^{\lambda-1} \sum_{j=1}^{k} \pi_j N(x_i^{(\lambda)} | \mu_j, \sigma_j^2),$$

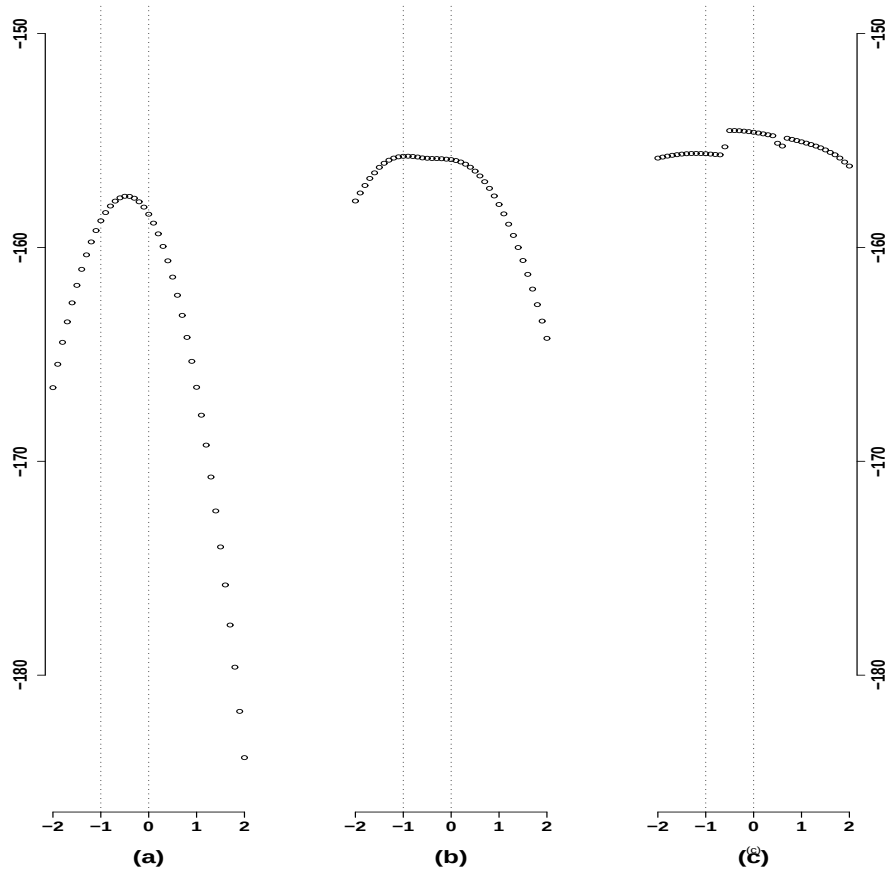

Figure 5: Likelihood of different power transformation conditional on (a)$k = 1$ (b)$k = 2$ and (c)$k = 3$.

where λ is the power transformation parameter and

$$x^{(\lambda)} = \begin{cases} (x^\lambda - 1)/\lambda & \text{if } \lambda \neq 0 \\ \log x & \text{if } \lambda = 0 \end{cases},$$

thus the Jacobian $x_i^{\lambda-1}$ is included in the likelihood.

Varying the power of the transformation from -2 to 2, where log transformation is already included, Figure 5 displays the corresponding likelihood conditional on $k = 1,\ 2,\ 3$. The inverse of square root transformation is suggested conditional on $k = 1$ and inverse transformation looks best when conditional on $k = 2$.

Table 1: Posterior probability of k, $P(k|x)$, for a simulated data set from $SN(3, 2, 5)$ over different models and different transformations: first column of $SN(x)$ for skew-normal model on the original scale, second column of $N(x)$ for normal model on the original scale, and third column for normal model on the transformation of power = -0.5.

k	SN (x)	N (x)	N (power = -0.5)
1	.8227	.0118	.6624
2	.1235	.2582	.2094
3	.0300	.3319	.0744
4	.0117	.2028	.0310
5	.0053	.0997	.0121
6	.0029	.0523	.0055
7	.0022	.0246	.0028
8	.0011	.0114	.0012
9	.0004	.0046	.0006
10	.0001	.0027	.0006

$k = 3$ results least amount of change in the likelihood but inverse of square root is plausible. Figure 5, therefore, shows the scale of the observation varies conditional on the number of components. Table 1 displays the posterior probability $P(k = j|Data),\ j = 1, ..., 10,$ for different models and scales. The probabilities in the first two columns use the original scale of x denoted by (x) and the last column is the results using the inverse of square root transformation. For the normal mixture model, the program Nmix is used which is developed by Peter Green and freely available at http://www.stat.bris.ac.uk/ peter implementing the reversible jump MCMC algorithm of mixture of normal distributions.

SN denotes skew-normal mixture model and N denotes normal mixture model. The results from first column and third column are pretty close in the terms of the posterior probability of the number of component, that is, both models attain the maximum posterior probability at $k = 1$ and relatively negligible amount at the others. The normal mixture model on the original scale attains its maximum posterior probability at $k = 3$. The relatively large amount of probability fairly spread over

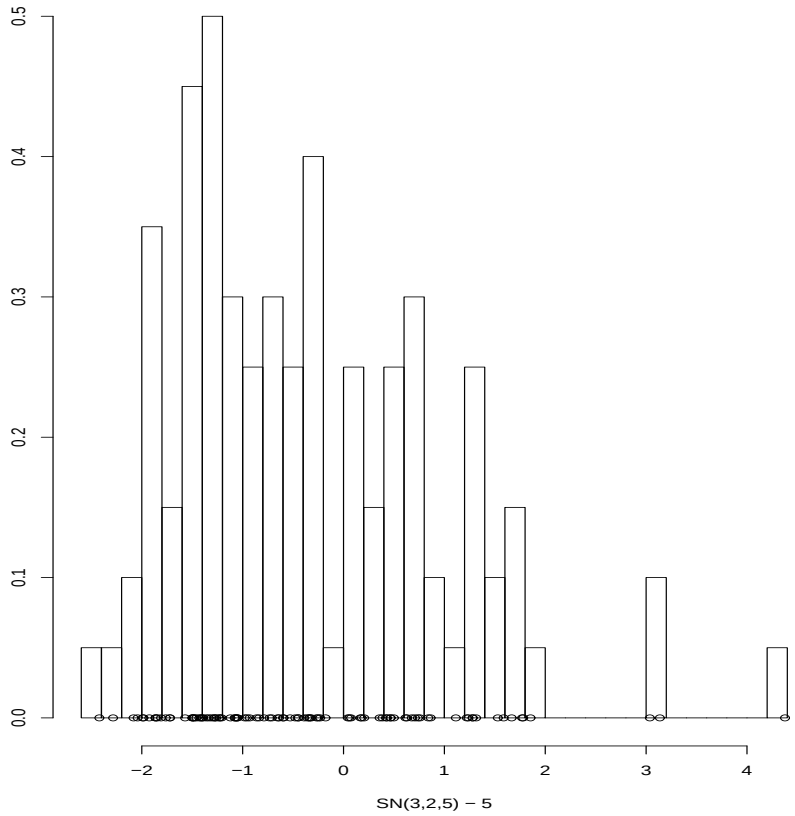

Figure 6: Histogram of 100 samples from SN(3,2,5)-5.

the large number of components. When we apply the normal assumption on the original scale, therefore, the spurious components are highly likely to arise. It would be appropriate to consider using power transformation or more flexible and wider class of distribution as the skew-normal family.

The above example shows how the transformation works without negative values from the observation.

Next, we add -5 to all the observation, which is equivalent to be sampled from $SN(-2, 2, 5)$. Figure 6 displays the histogram of the data after subtract 5 from the above data set. To make them all positive, we need to add a certain amount to the

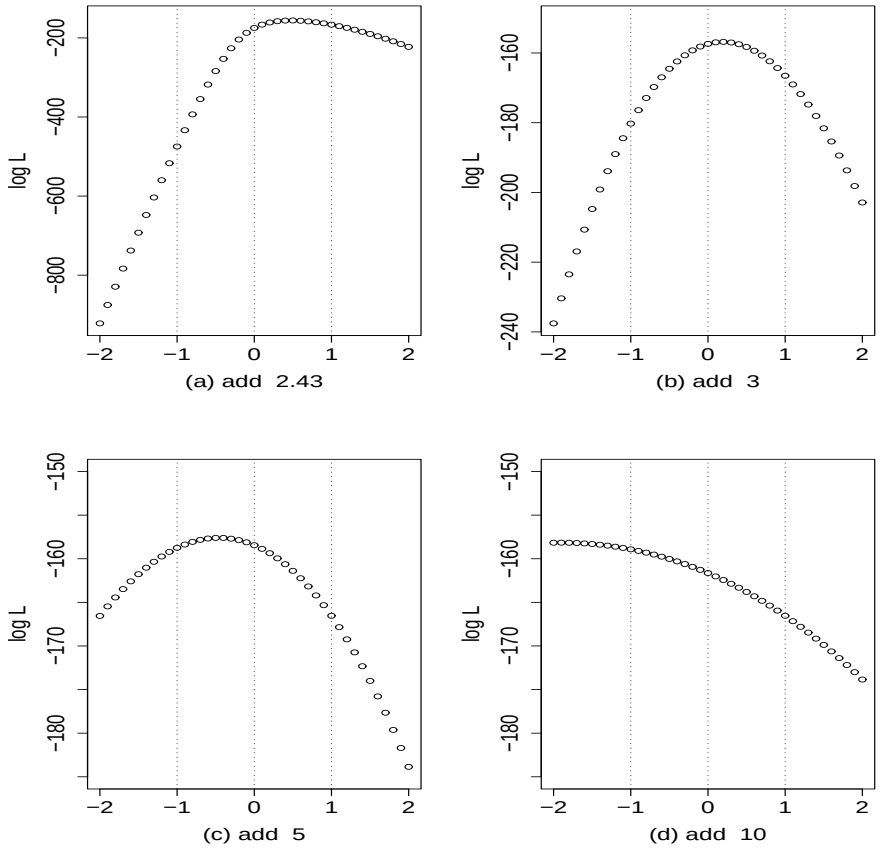

Figure 7: Log likelihood conditional on k of $SN(3, 2, 5) - 5$ after some amount of shift to make them positive.

data but do not know how large should it be. Schork and Schork (1988) showed the amount of such shift may affect the estimation of other parameters. They suggest to use cautiously pre-chosen amount before the analysis. Here we add different constants 2.43, 3, 5, 10.

Figure 7 shows the log likelihood conditional on $k = 1$ with different amount of shift. As it varies, the appropriate power of transformation varies, which makes it difficult the interpretation of the scale change. The parameter of power transformation could be treated as being indicative of how much the data lies far from normality.

Table 2: Posterior probability of k up to 10 from two different models for the tomato data set.

k	SN	normal
1	.5004	.0040
2	.2857	.3668
3	.1184	.2713
4	.0510	.1583
5	.0227	.0891
6	.0111	.0497
7	.0055	.0285
8	.0029	.0167
9	.0015	.0099
10	.0007	.0056

In such case, it may not be a matter that the power changes as the amount of shift varies. Even such case, sometimes we are required to make an interpretation of the transformation. The choice of the amount, then, could be problematic. The skew-normal family can encompass such abnormality into the model parameter and analyze the data on what we originally measured.

2.4.2 Tomato data set

We apply the same method to the data set analyzed in Gutierrez et al. (1995) denoted by Tomato. They adopted this mixture model of transformed normal components in an attempt to identify the number of underlying physical phenomena behind tomato root initiation. The observation y_i corresponds to the inverse proportion of the jth lateral root which expresses GUS ($j = 1, ..., 40$). This measurement is a possible indicator of the number of initial cells in the lateral root.

Table 2 displays the posterior probability via the reversible jump Markov chain Monte Carlo algorithm from both mixture of normals and skew-normals. We put the uniform discrete prior on the number of component, k, up to 15.

Table 3 shows the estimates of the parameters for each of the models in hand.

Table 3: Parameter estimates from 2 different models of the tomato data set.

parameter	one SN	two normals
weight	1	.76 – .24
location	1.37	1.95 – .45
scale	1.12	.43 – .82
skewness	6.79	–

The numbers in the parentheses indicate the estimate of the first component for the left, and then the estimate of the second component.

Figure 8 shows the MCMC samples of the number of component k for two different underlying distributions, (a) and (b) one for normal by the program Nmix by P. Green and (c) and (d) the other for skew-normal. Both plots show that the chains are mixing well. The plots of (c) and (d) show the convergence of both chains but slow convergence of skew-normal model.

Figure 9 displays the observations in the x-axis and its histogram, the predictive density from one component skew-normal distribution (solid line) and mixture of two normal distributions (dashed line). The number of parameters are, of course, greater in 2 normals than that in one skew-normal so that two normals look much closer to the histogram than a skew-normal does.

With the MCMC samples, we can obtain the classification probability $P(z_i = j|x_i)$. Figure 10 shows those probabilities of two different model, 2 SN and 2 normals just for comparison purpose.

2.4.3 Enzyme data set

The final example is the data set first analyzed by Bechtel et al. (1993) denoted by Enzyme. The interest here is in identifying subgroups of slow of fast metabolizers as a marker of genetic polymorphism in the general population. Bechtel et al. (1993) identified a mixture of two skewed distributions by using maximum likelihood

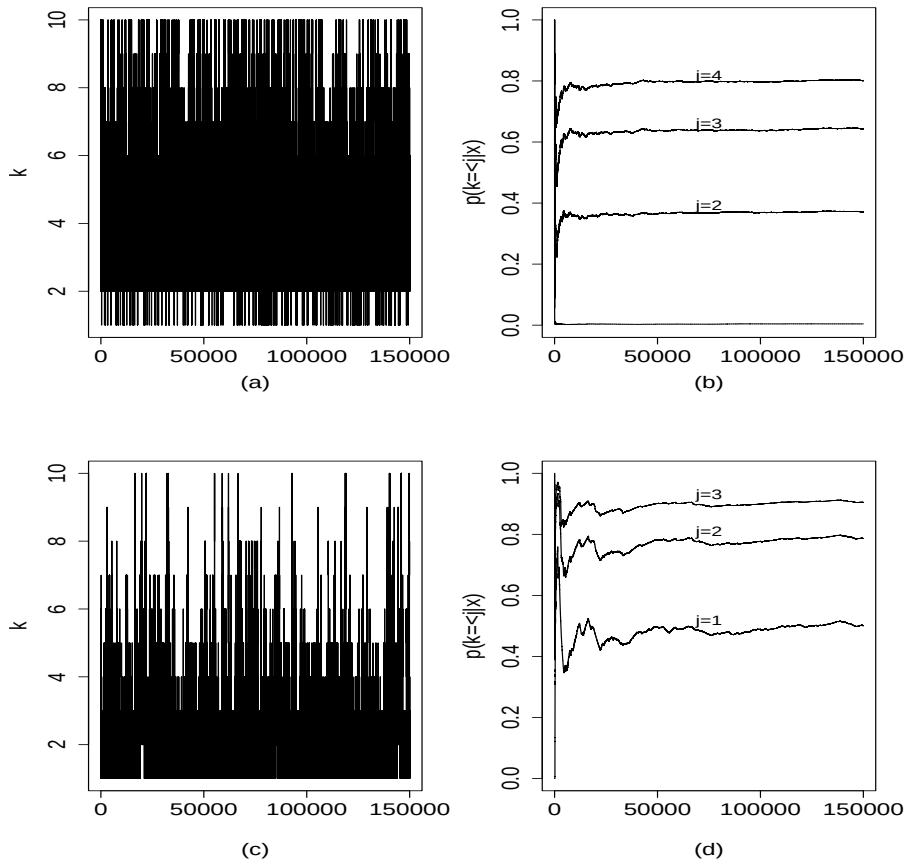

Figure 8: Plots of a trace of k and cumulative fraction (posterior probability of k) for the tomato data set, for 150 000 sweeps after 150 000 burn-in: (a) and (b) Normal mixture by Green's Nmix program and (c) and (d) Skew-normal mixture.

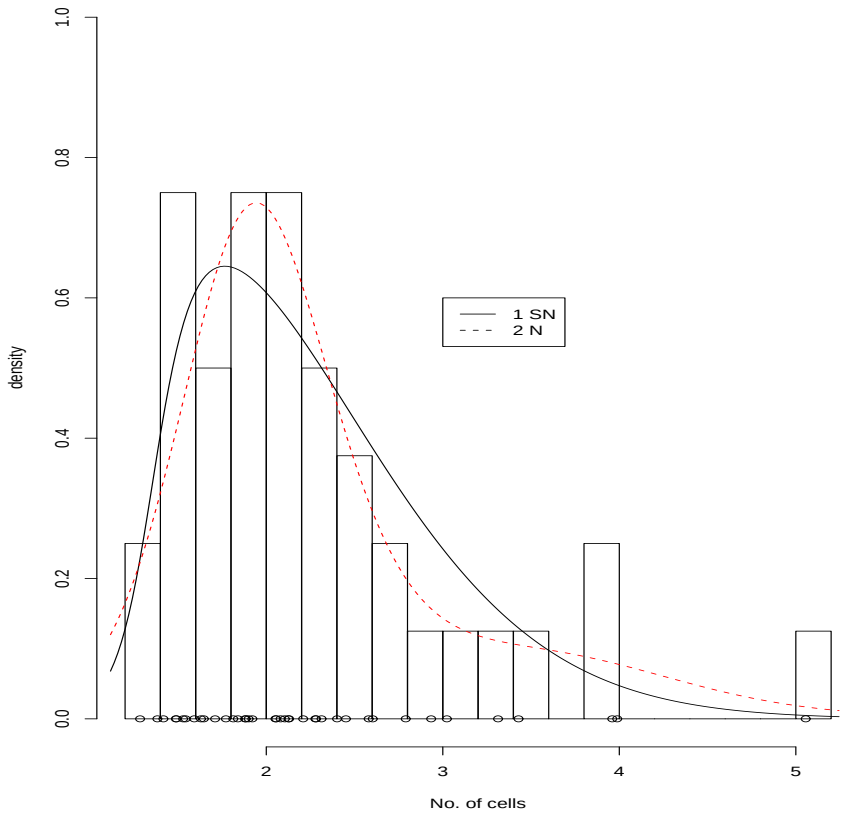

Figure 9: Predictive density in the different models: skew normal with $k = 1$ (solid line), and two components of normal (dotted line).

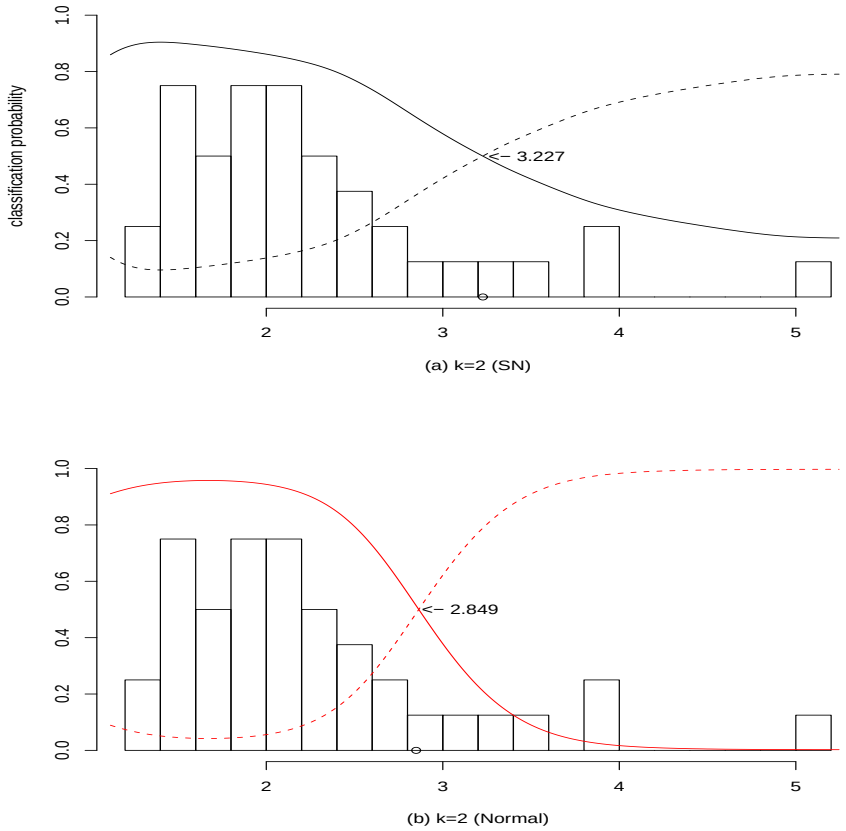

Figure 10: Plots of classification probability $P(s_i = j|x_i)$. Numbers in each plot denote the point where probabilities of different groups meet: (a) mixture of 2 SN (b) mixture of 2 normals.

Table 4: Posterior probability of k up to 10 from two different models for the enzyme data set.

k	SN	normal
1	.0006	0
2	.5598	.0282
3	.3018	.2863
4	.0996	.3120
5	.0284	.2069
6	.0074	.0978
7	.0016	.0407
8	.0005	.0175
9	.0002	.0064
10	.0000	.0027

Table 5: Parameter estimates from 3 different models of the enzyme data set.

parameter	two SNs	two normals	3 normals
weight	.62 – .38	.59 – .41	.60 – .20 – .20
location	.09 – .79	.19 – 1.26	.19 – 1.05 – 1.63
scale	.14 – .69	.08 – .51	.08 – .21 – .48
skewness	3.46 – 5.44	–	–

techniques developed by MacLean et al. (1976).

Table 4 displays a part of the posterior probability via the reversible jump Markov chain Monte Carlo algorithm from both mixture of normals and skew-normals. We put the uniform discrete prior on the number of component, k, up to 30.

Table 5 shows the estimates of the parameters for each of the models.

Figure 11 shows the MCMC samples of the number of component k for two different underlying distributions and the cumulative fraction of k. The results using normal mixture are displayed in (a) and (b), which are reproduced using the program Nmix by Green. Skew-normal mixture results are shown in (c) and (d).

The plots (b) and (d) show each chain remains stable indicating it converges. It is noticed that the model of skew-normal mixtures takes a little longer than normal mixture one.

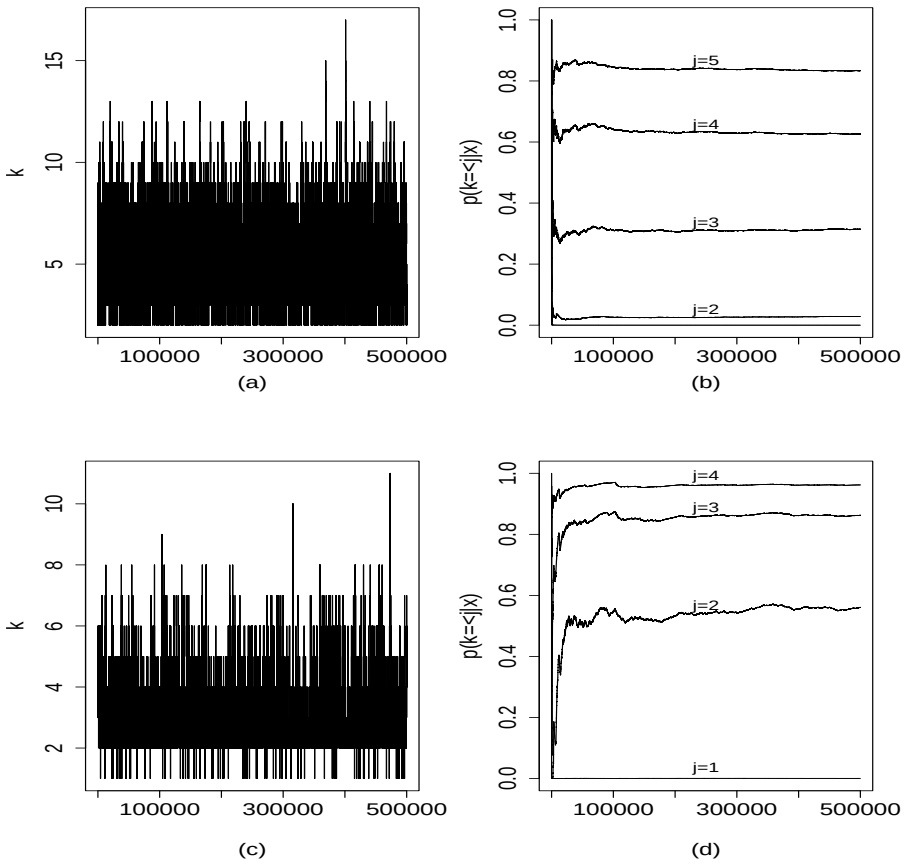

Figure 11: Plots of a trace of k and cumulative fraction (posterior probability of k) for the enzyme data set, for 500 000 sweeps after 500 000 burn-in: (a) and (b) Normal mixture by Green's Nmix program and (c) and (d) Skew-normal mixture.

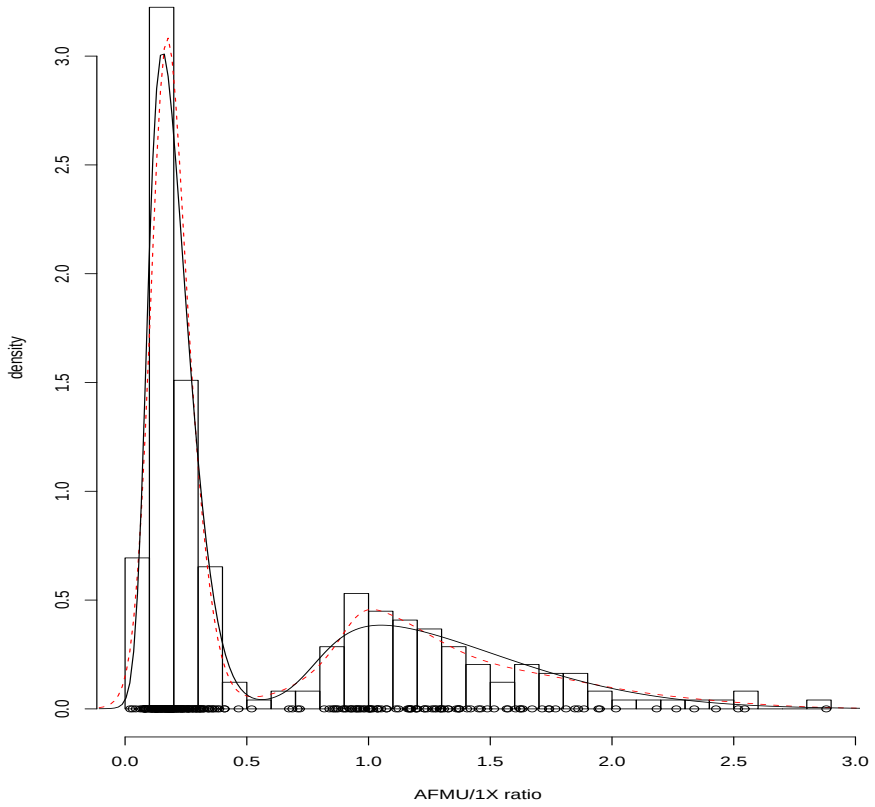

Figure 12: Estimated density in the different models: skew normal with $k = 2$ (solid line), and four components of normal (dotted line).

Figure 12 displays the observations on the x-axis and its histogram, the estimated density of the best models in terms of the posterior probability of k. Considering the fact that the number of parameters retained in each model is 7 for 2 skew-normal mixture and 11 for 4 normal mixture, the skew-normal mixture with smaller number of parameters fits the data as well as normal mixture. For the reference, Figure 13 shows the comparison of other models with different number of components in normal mixture.

With the MCMC samples, we can obtain the classification probability $P(z_i = j|x_i)$. Figure 14 shows those probabilities of two different model, 2 SN and 3 normals

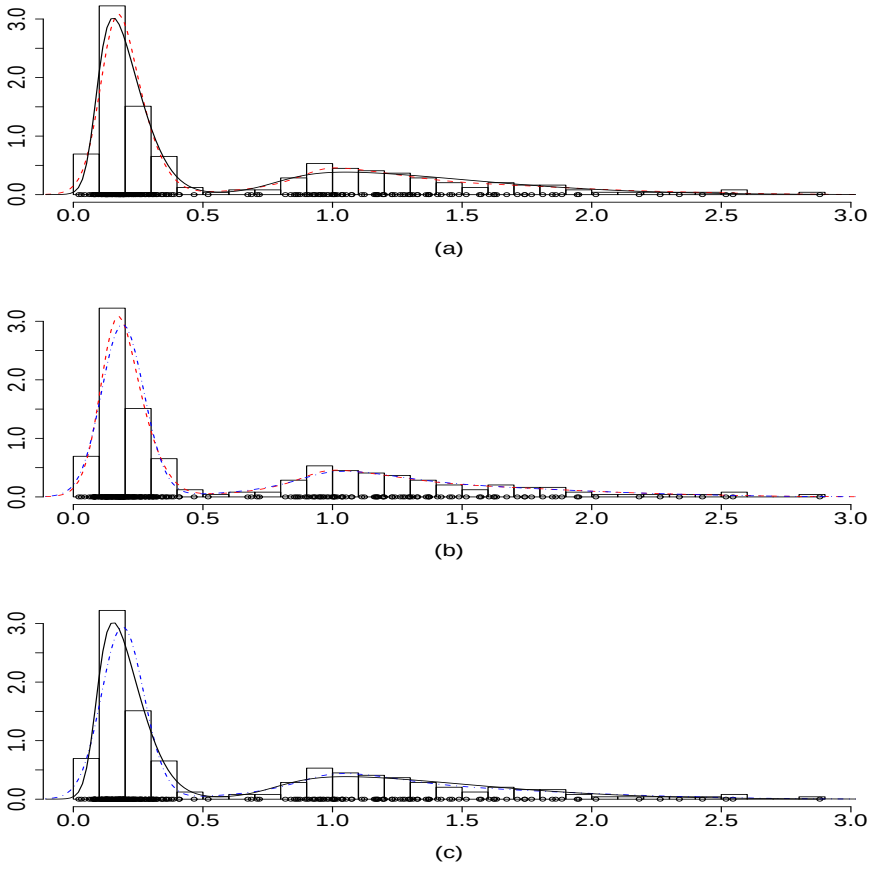

Figure 13: Estimated density in the different models: (a) 2 SN (solid line) vs 4 N (dashed line), (b) 4 N (dashed line) vs 3 N (dashed-dotted line) and (c) 2 SN (solid line) vs 3 N (dotted-dashed line).

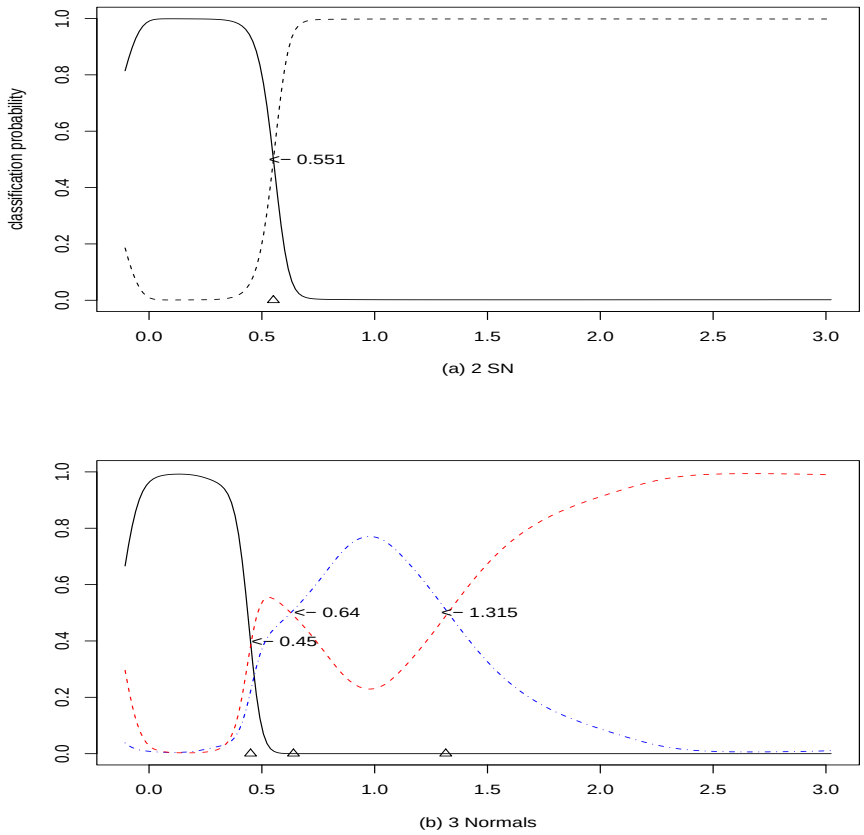

Figure 14: Plots of classification probability $P(s_i = j|x_i)$. Numbers in each plot denote the point where probabilities of different groups meet: (a) mixture of 2 SN (b) mixture of 3 normals.

since 3 normal mixture is more interpretable than 4 normal mixture.

Figure 15 also displays the classification probability with different number of group, 3 SN and 3 normals. The huge variance of third component of the normal model makes strange group allocation in the middle of the data, which is not observed in the skew-normal mixture model. The group allocation, therefore, will be nicely done using skew-normal mixture model.

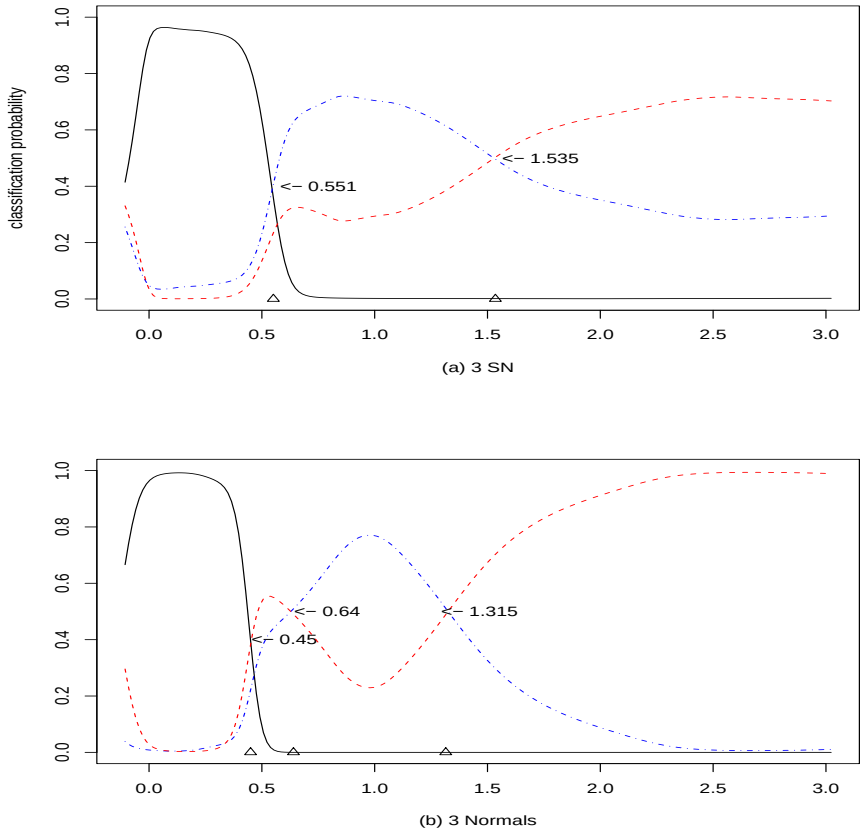

Figure 15: Plots of classification probability $P(s_i = j|x_i)$. Numbers in each plot denote the point where probabilities of different groups meet: (a) mixture of 3 SN (b) mixture of 3 normals.

2.5 Discussion

We have proposed a mixture of skew-normal distribution as an alternative to change of the original scale including the Box-Cox transformation when there exists a certain amount of skewness in the data. The number of components is estimated by the reversible jump MCMC algorithm which reports the posterior probability of the number of components. When there are negative values, the Box-Cox transformation needs another parameter to shift the data to make them all positive. The amount of shift has been reported to affect the entire estimation procedure and the inference of the number of components. In application to some real data and simulated data, the mixture of skew-normal distribution have drawn same conclusion as that obtained from transformation without changing the scale, which produces much easier interpretation of the results. The following subsections consider the prior on the skewness parameter in the skew-normal distribution and the integration method for future research.

2.5.1 Integration

When the conjugacy is acquired by the prior structure, then the marginal likelihood

$$p(Y|k) = \frac{p(\theta_k|k)p(Y|k,\theta_k)}{p(\theta_k|k,Y)}$$

can be calculated and the posterior probabilities $p(k|Y)$ can be also obtained since it is proportional to $p(k)p(Y|k)$. In such a case, Bayesian inference about k and θ_k given k could be conducted separately.

There is difficulty in attaining the conjugacy for the skew-normal distribution. As described in this article, it could not be problematic in the univariate case. When the extension to the multivariate case is considered, the dimension reduction of the unknowns by integrating out a part of the set of parameters could help implement

the reversible jump MCMC algorithm. In the mixture of multivariate skew-normal distributions, the scale and skewness parameters can be integrated out to reduce much of the dimensions to ease the dimension varying MCMC.

2.5.2 Moment matching condition

There are a few other schemes developed to incorporate the dimension varying situation, for example, Stephens (2000) and Carlin and Chib (1995) among which we are solely dependent on the algorithm by Green (1995) It could be challenging to use other algorithms to move across the models with different number of components in the mixture model. Also, within the reversible jump algorithm, we can take different approach for the update of the skewness parameter in the moment matching condition or it is possible to consider more efficient proposal choice, for example, Brooks et al. (Brooks et al., 2003) .

2.5.3 Prior on the skewness parameter

The normal distribution with mean zero has been taken as the prior for the skewness parameter λ providing that there exists a certain amount of skewness in the data. As we put another hierarchical level of prior for the scale parameter, a prior will be taken for the hyperparameter b_3.

Liseo and Loperfido (2005) showed that the Jeffreys' prior of the skewness parameter in the skew-normal distribution is proper, which could be considered for the prior distribution for λ.

CHAPTER III

THE MULTIVARIATE SKEW-NORMAL MIXTURE

3.1 Introduction

Suppose that $X = (X_1, ...X_d)^\top$ is a d-dimensional multivariate random variable. A multivariate extension of the Box-Cox transformation is generalized in the k component of mixture context by Schork and Schork (1988) as a multivariate generalization of the technique of MacLean et al. (1976). Such transformation is applied to each variable as

$$y_m = \begin{cases} (x_m^{\lambda_m} - 1)/\lambda_m + \lambda_m & \text{if } \lambda \neq 0 \\ \log(x_m) & \text{if } \lambda = 0 \end{cases},$$

where y_m is the transformed and x_m the untransformed variable ($m = 1, ..., d$). As noted in the previous chapter, the variables should be translated to make the data all positive before analysis in the presence of negative values. To allow such location parameter to vary in a numerical routine increases the instability of convergence while adding an unimportant parameter to be estimated (Schork and Schork, 1988). Thus, Schork and Schork (1988) suggested that the value be established with discretion by the user before analysis. We observed in the previous chapter that the amount of such translation has an effect on the inference of the mixture model, especially when the number of components is unknown.

While MacLean et al. (1976) and Schork and Schork (1988) modeled the skewness in the mixture context, McLachlan et al. (2002) and Peel and McLachlan (2000) considered the multivariate t-distribution to incorporate heavy tails along with mul-

tiple modes.

Azzalini and Capitanio (1999) pointed out the possible problems when the transformation of the variables is taken into account to achieve multivariate normality, these are

(a) the transformations are usually on each component separately, and achievement of joint normality is only hoped for;

(b) the transformed variables are more difficult to interpret, especially when each variable is transformed by using a different function; and

(c) when multivariate homoscedasticity is required, this often requires a transformation which is different from the transformation for normality.

As an alternative to the Box-Cox transformation, we suggest a mixture model approach using the multivariate skew-normal distribution.

Azzalini and Capitanio (1999) demonstrate that the multivariate skew normal distribution has reasonable flexibility in real data fitting, while it maintains some convenient formal properties of the normal density. Discriminant analysis was presented as an application of multivariate skew normal distribution in their paper.

In this chapter, we present a methodology for model-based clustering using the mixture of multivariate skew-normal distributions. In section 3.2, a Bayesian estimation procedure for the multivariate skew-normal distribution is considered, and it is extended to the mixture model in section 3.3 followed by some applications in section 3.4.

3.2 Multivariate Skew-normal Distribution

A multivariate version of the skew-normal distribution was formalized in Azzalini and Dalla Valle (1996) and Azzalini and Capitanio (1999). A d-dimensional random

variable Z is said to have a multivariate skew normal distribution if it is continuous with density function, denoted by $SN_d(\Omega, \alpha)$

$$2\phi_d(z;\Omega)\Phi(\alpha^\top z), \quad z \in R^d,$$

where $\phi_d(z;\Omega)$ is the d-dimensional normal density with zero mean and correlation matrix Ω, $\Phi(\cdot)$ is the $N(0,1)$ distribution function and α is a d-dimensional vector. The matrix Ω is called the scale parameter and α is called the shape or skewness parameter. It is an extended version of normal family since when $\alpha = 0$, the density reduces to $N_d(0,\Omega)$. Another representation of skew-normal distribution is useful for the Bayesian estimation. Suppose that

$$\begin{pmatrix} U \\ X \end{pmatrix} \sim 2 \cdot I(U > 0) \cdot \ N_{d+1}(\mathbf{0}, \Omega^*), \quad \Omega^* = \begin{pmatrix} 1 & \delta^\top \\ \delta & \Omega \end{pmatrix}$$

where U is a scalar component, X and δ are vectors of size d and Ω^* is a correlation matrix. Then the marginal distribution of X is $SN_d(\Omega, \alpha)$ where

$$\alpha = \frac{1}{(1 - \delta^\top \Omega^{-1} \delta)^{1/2}} \Omega^{-1} \delta.$$

The multivariate skew-normal distribution can also be expressed using a covariance matrix. Suppose that

$$\begin{pmatrix} U \\ X \end{pmatrix} \sim 2 \cdot I(U > a_0) \cdot \ N_{d+1}(\boldsymbol{\mu}, \Sigma), \quad \Sigma = \begin{pmatrix} \sigma_0^2 & \Sigma_{21}^\top \\ \Sigma_{21} & \Omega \end{pmatrix},$$

where $\mu = (a_0, \xi^\top)^\top$. Then, the marginal distribution of X is

$$2\phi_d(x - \xi; \Omega)\Phi(\alpha^\top \omega^{-1}(x - \xi))$$

where ω is the diagonal matrix of square root of diagonal elements of Ω and

$$\alpha^\top \omega^{-1}(x - \xi) = \frac{\Sigma_{21}^\top \Omega^{-1}(x - \xi)}{\sqrt{\sigma_0^2 - \Sigma_{21}^\top \Omega^{-1} \Sigma_{21}}}. \tag{3.1}$$

Thus,

$$\alpha = \frac{\omega\Omega^{-1}\Sigma_{21}}{\sqrt{\sigma_0^2 - \Sigma_{21}^\top\Omega^{-1}\Sigma_{21}}}.$$

The marginal distribution of X is not dependent on the mean parameter of U, a_0, but it seems to depend on σ_0^2, the variance of the integrated variable U. Dividing by σ_0 both denominator and numerator, however, shows that the marginal distribution of X does not depend on σ_0^2. From the truncated multivariate normal distribution, we know that the conditional distribution of U given X is a truncated normal distribution with the support of (a_0, ∞) proportional to

$$I(U > a_0)N(U|a_0 + \Sigma_{21}^\top\Omega^{-1}(X - \xi), \sigma_0^2 - \Sigma_{21}^\top\Omega^{-1}\Sigma_{21}),$$

and the normalizing constant is

$$1 - \Phi\left[\Sigma_{21}^\top\Omega^{-1}(X - \xi)/\sqrt{\sigma_0^2 - \Sigma_{21}^\top\Omega^{-1}\Sigma_{21}}\right].$$

Now, a_0 and σ_0^2 are also included in the set of parameters for the ease of computation, thus the set of all the unknown parameters includes $\mu = (a_0, \xi^\top)^\top$ and Σ, instead of just ξ, Ω and α. The estimation procedure on the multivariate skew-normal distribution is now similar to the EM structure. Given a multivariate skew-normal distribution, we can induce a truncated univariate normal distribution treated as a missing quantity. Both yields a truncated multivariate normal distribution for their joint distribution.

The priors that we need in Bayesian procedure for the parameters μ and Σ are conjugate priors, these are

$$\Sigma^{-1} \sim Wish(r_p, W_p), \text{ and}$$

$$\mu = (a_0, \xi^\top)^\top \sim N_{d+1}(\mu_p, \Sigma/\kappa_p),$$

where r_p, Wishart degrees of freedom, is positive scalar and W_p, Wishart scale matrix, is a $(d+1) \times (d+1)$ matrix.

Let $x_1, ..., x_n$ be from $SN(\xi, \Omega, \alpha)$, $u_1, ..., u_n$ are missing and $y_i = (u_i, x_i^\top)^\top$. Then, the full conditional distributions, which are necessary for the Gibbs sampler to explore the posterior distribution, are as follows:

$$p(u_i|\cdots) \propto I(u_i > a_0) \cdot N(a_0 + \Sigma_{21}^\top \Omega^{-1}(x_i - \xi),\ \sigma_0^2 - \Sigma_{21}^\top \Omega^{-1} \Sigma_{21})$$

$$p(\mu|\cdots) \propto I(a_0 < \min_i u_i) \cdot N_{d+1}\left(\frac{n\bar{y} + \kappa_p \mu_p}{n + \kappa_p}, \frac{\Sigma}{n + \kappa_p}\right)$$

and

$$p(\Sigma^{-1}|\cdots) \sim Wish(n + r_p, [W_p^{-1} + \sum (y_i - \mu)(y_i - \mu)^\top + \kappa_p(\mu - \mu_p)(\mu - \mu_p)^{-1}]^{-1}),$$

which yields the following estimation procedure.

Let $x_1, ..., x_n$ be from $SN_d(\xi, \Omega, \alpha)$. Then, given the observations $x_1, ..., x_n$ the Gibbs sampler can be implemented in the following way at each iteration step and the parameter estimation is based on these samples at each iteration:

(a) Generate u_i from a truncated normal distribution

$$N(a_0 + \Sigma_{21}^\top \Omega^{-1}(x_i - \xi),\ \sigma_0^2 - \Sigma_{21}^\top \Omega^{-1} \Sigma_{21}) \cdot\ I(u_i > a_0)$$

and let $y_i = (u_i, x_i^\top)^\top$.

(b) Generate a sample of μ from a truncated multivariate normal distribution

$$N_{d+1}\left(\frac{n\bar{y} + \kappa_p \mu_p}{n + \kappa_p}, \frac{\Sigma}{n + \kappa_p}\right) \cdot I(a_0 < \min_i u_i)$$

(c) Generate a sample of Σ^{-1} from a Wishart distribution

$$inv - Wish(n + r_p, [W_p^{-1} + \sum (y_i - \mu)(y_i - \mu)^\top + \kappa_p(\mu - \mu_p)(\mu - \mu_p)^{-1}]^{-1}).$$

3.3 Mixture of Multivariate Skew-normal Model

Throughout this section, it is assumed that the number of components is known. The $d \times 1$ random vector X is said to follow a mixture of k multivariate $SN_d(\xi_j, \Omega_j, \alpha_j)$ if it has the density function

$$f(x) = \sum_{j=1}^{k} w_j SN_d(x|\xi_j, \Omega_j, \alpha_j) = \sum_{j=1}^{k} w_j 2 \cdot |2\pi\ \Omega_j|^{-.5} e^{-\frac{1}{2}(x-\xi_j)^\top \Omega_j^{-1}(x-\xi_j)} \Phi(\alpha_j^\top \omega_j^{-1}(x-\xi_j)),$$

where $\boldsymbol{w} = (w_1, ..., w_k)^\top,\ \boldsymbol{\xi} = (\xi_1, ..., \xi_j),\ \boldsymbol{\Omega} = (\Omega_1, ..., \Omega_k),\ \boldsymbol{\alpha} = (\alpha_1, ..., \alpha_k),\ w_j \in (0, 1),\ \xi_j \in R^d,\ \alpha_j \in R^d$, Ω_j is a $d \times d$ matrix, and ω_j is the diagonal matrix with lth diagonal elements equal to the square root of lth diagonal elements of Ω_j. Define the new set of parameters μ_j and Σ_j for $j = 1, ..., k$ by

$$\mu_j = \begin{pmatrix} a_{j,0} \\ \xi_j \end{pmatrix}, \quad \Sigma_j = \begin{pmatrix} \sigma_{j,0}^2 & \Sigma_{j,21}^\top \\ \Sigma_{j,21} & \Omega_j \end{pmatrix},$$

from which we know that

$$\alpha_j^\top \omega_j^{-1}(x - \xi_j) = \frac{\Sigma_{j,21}^\top \Omega_j^{-1}(x - \xi_j)}{\sqrt{\sigma_{j,0}^2 - \Sigma_{j,21}^\top \Omega_j^{-1} \Sigma_{j,21}}}$$

and

$$\alpha_j = \frac{\omega_j \Omega_j^{-1} \Sigma_{j,21}}{\sqrt{\sigma_{j,0}^2 - \Sigma_{j,21}^\top \Omega_j^{-1} \Sigma_{j,21}}}.$$

The following priors can then be used:

$$\begin{aligned} \boldsymbol{w} &\sim Dirichlet(a, ..., a) \\ \beta &\sim Wishart_{d+1}(2g_p, (2h_p)^{-1}) \\ \tau_j &\sim Wishart_{d+1}(2r_p, (2\beta)^{-1}) \\ \mu_j &\sim N_{d+1}(\mu_p, \kappa_p^{-1}), \end{aligned}$$

where $\tau_j = \Sigma_j^{-1}$. Note that they are all $(d+1) \times (d+1)$ matrices except μ_j which is a $d+1$ vector. As in Stephens (1997), we allow another level of prior on the

hyperparameter of the precision matrix τ_j. Note that when $W \sim Wishart(v, S)$, the probability density function of W, $p(W)$, is

$$p(W) \propto |S|^{-v/2}|W|^{(v-d-1)/2}e^{-\frac{1}{2}tr(S^{-1}W)}.$$

We choose $g_p = 0.3,\ h_p = 3$ and $a = 1$ as in Stephens (1997) and the others are based on the range of the observed data as in Richardson and Green (1997) except B_3 which reflects the belief of a researcher about how far the data would be from normality.

The group allocation variable $\boldsymbol{z}^n$ is a set of missing observation as in the Chapter II that is necessary for the Gibbs sampler. The unobservable variable u_i also needs to be generated in the Gibbs algorithm so that the joint distribution of u_i and x_i becomes the truncated $d+1$-dimensional multivariate normal distribution with parameters μ_j and Σ_j. Therefore, the joint distribution of all the unknown parameters including $\boldsymbol{z}^n$ and $\boldsymbol{u}^n$ for fixed k becomes

$$\begin{aligned} p(\boldsymbol{w}, \beta, \boldsymbol{\tau}, \boldsymbol{\xi}, \boldsymbol{\psi}, \boldsymbol{x}^n, \boldsymbol{u}^n, \boldsymbol{z}^n) \quad \propto \quad & |\beta|^{(2g_p-d-1)/2}e^{-\frac{1}{2}tr(2h_p\beta)} \\ & \prod_{j=1}^{k}\Big[w_k^a e^{-\frac{1}{2}(\mu_j-\mu_p)^\top \kappa_p(\mu_j-\mu_p)}|2\beta|^{r_p}|\tau_j|^{(2r_p-d-1)/2}e^{-\frac{1}{2}tr(2\beta\tau_j)} \\ & \prod_{i\in N_j} w_j MVN_{d+1}(y_i|\mu_j, \Sigma_j)\cdot I(u_i > a_{j,0})\Big], \end{aligned}$$

where $y_j = (u_j, x_j^\top)^\top$, a vector of size $d+1$. The posterior distribution, which is proportional to the joint distribution, is simply written as

$$p(\boldsymbol{z}^n, \boldsymbol{u}^n, \boldsymbol{w}, \boldsymbol{\xi}, \boldsymbol{\tau}, \beta|\boldsymbol{x}^n, k) \propto \prod_{j=1}^{k}\prod_{i\in N_j} w_j MVN_{d+1}(y_i|\mu_j, \Sigma_j)p(\boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{\Sigma})\cdot I(u_i > a_{j,0}).$$

Multivariate normal theory shows that

$$\begin{aligned} MVN_{d+1}(y_i|\mu_j, \Sigma_j)\cdot I(u_i > a_{j,0}) \quad = \quad & MVN_d(x_i|\xi_j, \Omega_j)\cdot I(u_i > a_{j,0})\cdot \\ & N(u_i|a_{j,0} + \Sigma_{j,21}^\top \Omega_j^{-1}(x_i - \xi_j),\ \sigma_{j,0}^2 - \Sigma_{j,21}^\top \Omega_j^{-1}\Sigma_{j,21}). \end{aligned}$$

The full conditional distributions are then obtained as follows:

$$\begin{aligned}
\boldsymbol{w}|\cdots &\sim Dirichlet(a+n_1,...,a+n_k)\\
P(\boldsymbol{z}_{ij}=1|\cdots) &\propto w_j SN_d(x_i|\xi_j,\Omega_j,\alpha_j)\\
\beta|\cdots &\sim Wish_{d+1}(2g_p+2kr_p,(2h_p+2\sum_{j=1}^{k}\tau_j)^{-1})\\
u_i|\cdots &\sim I(u_i>a_{j,0})\cdot N(a_{j,0}+\Sigma_{j,21}^{\top}\Omega_j^{-1}(x_i-\xi_j),\ \sigma_{j,0}^2-\Sigma_{j,21}^{\top}\Omega_j^{-1}\Sigma_{j,21})\\
\mu_j|\cdots &\propto N_{d+1}((n_j\tau_j+\kappa_p)^{-1}(n_j\tau_j\bar{y}_j+\kappa_p\mu_p),(n_j\tau_j+\kappa_p)^{-1})\cdot I(a_{j,0}<\min_{i\in N_j}u_i)\\
\tau_j|\cdots &\sim Wish_{d+1}(2r_p+n_j,[2(\beta+0.5\Sigma_{i\in N_j}(y_i-\mu_j)(y_i-\mu_j)^{\top})]^{-1}).
\end{aligned}$$

Let

$$\phi_j=(\phi_{j,1},\phi_{j,2}^{\top})^{\top}=(n_j\tau_j+\kappa_p)^{-1}(n_j\tau_j\bar{y}_j+\kappa_p\mu_p)$$

and

$$A_j=\begin{pmatrix} A_{j,11} & A_{j,12}\\ A_{j,21} & A_{j,22}\end{pmatrix}=(n_j\tau_j+\kappa_p)^{-1}.$$

Applying the multivariate theory again, we obtain

$$\xi_j|\cdots\sim MVN_d(\phi_{j,2},A_{j,22})$$

and

$$a_{j,0}|\cdots\sim I(a_{j,0}<\min_{i\in N_j}u_i)\cdot N(\phi_{j,1}+A_{j,12}A_{22,j}^{-1}(\xi_j-\phi_{j,2}),A_{j,11}-A_{j,12}A_{j,22}^{-1}A_{j,21}).$$

Given the observations $x_1,...,x_n$, the Gibbs algorithm for the update can be implemented in the following way :

(a) Update the allocation vector $\boldsymbol{z}_i$ from

$$P(z_{ij}=1|\cdots)\propto w_j SN_d(x_i|\xi_j,\Omega_j,\alpha_j)$$

(b) Update the parameter β from

$$p(\beta|\cdots) \sim Wish_{d+1}(2g_p + 2kr_p, (2h_p + 2\sum_{j=1}^{k}\tau_j)^{-1})$$

(c) Update the missing u_is for $j = 1, ..., k$ from

$$p(u_i|\cdots) \sim I(u_i > a_{j,0}) \cdot N(a_{j,0} + \Sigma_{j,21}^{\top}\Omega_j^{-1}(x_i - \xi_j),\ \sigma_{j,0}^2 - \Sigma_{j,21}^{\top}\Omega_j^{-1}\Sigma_{j,21})$$

letting $y_i = (u_i, x_i^{\top})^{\top}$

(d) With $\bar{y}_j = \sum_{i \in N_j} y_i/n_j$, let

$$\phi_j = (\phi_{j,1}, \phi_{j,2}^{\top})^{\top} = (n_j\tau_j + \kappa_p)^{-1}(n_j\tau_j\bar{y}_j + \kappa_p\mu_p)$$

and

$$A_j = \begin{pmatrix} A_{j,11} & A_{j,12} \\ A_{j,21} & A_{j,22} \end{pmatrix} = (n_j\tau_j + \kappa_p)^{-1},$$

update ξ_j from

$$\xi_j|\cdots \sim MVN_d(\phi_{j,2}, A_{j,22})$$

and $a_{j,0}$ from

$$a_{j,0}|\cdots \sim I(a_{j,0} < \min_{i \in N_j} u_i) \cdot N(\phi_{j,1} + A_{j,12}A_{22,j}^{-1}(\xi_j - \phi_{j,2}), A_{j,11} - A_{j,12}A_{j,22}^{-1}A_{j,21}).$$

(e) Update τ_j from

$$\tau_j|\cdots \sim Wish_{d+1}(2r_p + n_j, [2(\beta + 0.5\Sigma_{i \in N_j}(y_i - \mu_j)(y_i - \mu_j)^{\top})]^{-1})$$

and let $\Sigma_j = \tau_j^{(-1)}$,

$$\sigma_{j,0} = \Sigma_{j,11}, \quad \Omega_j = \Sigma j, 22,$$

and

$$\alpha_j = \frac{\omega_j\Omega_j^{-1}\Sigma_{j,21}}{\sqrt{\sigma_{j,0}^2 - \Sigma_{j,21}^{\top}\Omega_j^{-1}\Sigma_{j,21}}}.$$

(f) Update $\boldsymbol{w}$ from

$$Dirichlet(a + n_1, ..., a + n_k).$$

3.4 Bayes Factor

3.4.1 Bayes factor

Let M_1 and M_2 denote two different models in interest, and take a look at the posterior odds ratio

$$\frac{p(M_1|y)}{p(M_2|y)} = \frac{p(y|M_1)}{p(y|M_2)} \times \frac{p(M_1)}{p(M_2)},$$

where y denotes the set of observations and $p(M_i)$ is the prior probability of $M_i,\ i = 1, 2$ being the true model. To determine this ratio we need to multiply the prior odds ratio by what is known as the marginal likelihood ratio (also known as the integrated likelihood ratio of prior predictive) (Denison et al., 2002). The marginal likelihood of model M_i gives a measure of the probability of observing the data given that M_i is true. To account for the uncertainty in the unknowns associated with each model we determine the marginal likelihood by integrating out the model parameters. The Bayes factor is defined for the comparison of two competing models and, if we wish to consider the relative merits of M_i over M_j, is given by

$$BF_{ij} \equiv (M_i, M_j) = \frac{p(M_i|D)}{p(M_j|D)} / \frac{p(M_i)}{p(M_j)},$$

the posterior to prior odds ratio. In the case where the prior probabilities of each model have been taken to be equal, we find that the Bayes factor is exactly the same as the posterior odds ratio. In this case, choosing the model with the highest posterior probability is equivalent to picking the model whose Bayes factor with respect to any other model is greater than one. Kass and Raftery (1995) suggest that, if the Bayes factor for M_i over M_j is between 1 and 3, then there is little perceived difference between the models, between 3 and 20 there is positive evidence in favor of M_i, 20 to 150 strong evidence and, if the Bayes factor is over 150, there is very strong evidence in favor of M_i. Under the equal model prior probabilities, Bayes factor reduces to the

ratio of marginal likelihood. In the following subsection we describe how to obtain the marginal likelihood of multivariate skew-normal mixture model using the method developed by Chib (1995).

3.4.2 Marginal likelihood of multivariate skew-normal mixture

Let M_k denote the skew-normal mixture model with k components with density function

$$p(\boldsymbol{x}^n|\boldsymbol{w},\boldsymbol{\theta}) = \prod_i^n \sum_{j=1}^k w_j MSN(x_i|\xi_j,\Omega_j,\alpha_j),$$

where $\boldsymbol{w} = (w_1,...,w_k)^\top$, $\boldsymbol{\theta} = (\theta_1,...,\theta_k)$ with $\theta_j = (\xi_j,\Omega_j,\alpha_j)$.

Let $\boldsymbol{\psi} = (\psi_1,...,\psi_k)$, and $\psi_j = (a_{j,0},\sigma^2_{j,0})$. Finally $\boldsymbol{\mu} = (\mu_1,...,\mu_k)$ and $\boldsymbol{\Sigma} = (\Sigma_1,...,\Sigma_k)$, where $\mu_j = (a_{j,0},\xi_j)$and $\Sigma_j = \begin{pmatrix} \sigma^2_{j,0} & \Sigma^\top_{j,21} \\ \Sigma_{j,21} & \Omega_j \end{pmatrix}$ with $\tau_j = \Sigma_j^{-1}$ or we can rewrite $(\boldsymbol{\mu},\boldsymbol{\Sigma}) = (\boldsymbol{\theta},\boldsymbol{\psi})$.

Then, the joint distribution of the data and the unknown parameters is

$$\begin{aligned} p(\boldsymbol{x}^n,\boldsymbol{w},\boldsymbol{\theta},\boldsymbol{\psi},\beta) &= p(\boldsymbol{x}^n,\boldsymbol{w},\boldsymbol{\mu},\boldsymbol{\tau},\beta) && (3.2)\\ &= p(\boldsymbol{x}^n|\boldsymbol{w},\boldsymbol{\mu},\boldsymbol{\tau},\beta)p(\boldsymbol{w},\boldsymbol{\mu},\boldsymbol{\tau},\beta) && (3.3)\\ &= p(\boldsymbol{w},\boldsymbol{\mu},\boldsymbol{\tau},\beta|\boldsymbol{x}^n)m(\boldsymbol{x}^n), && (3.4)\end{aligned}$$

from which we get the following equality

$$m(\boldsymbol{x}^n) = \frac{p(\boldsymbol{x}^n|\boldsymbol{w},\boldsymbol{\mu},\boldsymbol{\tau},\beta)p(\boldsymbol{w},\boldsymbol{\mu},\boldsymbol{\tau},\beta)}{p(\boldsymbol{w},\boldsymbol{\mu},\boldsymbol{\tau},\beta|\boldsymbol{x}^n)}.$$

The above equation is referred to the basic marginal likelihood identity(BMI) (Chib, 1995), which holds for any value of $\boldsymbol{\Xi} = (\boldsymbol{w},\boldsymbol{\mu},\boldsymbol{\tau},\beta)$. Chib (1995) showed an estimate of the marginal likelihood $m(\boldsymbol{x}^n)$ by estimating the posterior density when all complete conditional densities used in the Gibbs sampler have closed-form expressions, which is the case here. For a given $\boldsymbol{\Xi}$ (say $\boldsymbol{\Xi}^*$), the posterior ordinate $p(\boldsymbol{\Xi}^*|\boldsymbol{x}^n)$,

evaluated at Ξ^*, can be estimated by exploiting the information in the collection of complete conditional densities, denoted by $\hat{p}(\boldsymbol{\Xi}^*|\boldsymbol{x}^n)$. The proposed estimate becomes, on the log scale,

$$\log \hat{m}(\boldsymbol{x}^n) = \log p(\boldsymbol{x}^n|\boldsymbol{\Xi}^*) + \log p(\boldsymbol{\Xi}^*) - \log \hat{p}(\boldsymbol{\Xi}^*|\boldsymbol{x}^n).$$

The posterior density can be computed from the decomposition

$$\begin{aligned} p(\boldsymbol{\Xi}^*|\boldsymbol{x}^n) &= p(\boldsymbol{\xi}^*|\boldsymbol{x}^n) \times p(\boldsymbol{a}_0^*|\boldsymbol{\xi}^*, \boldsymbol{x}^n) \times p(\beta^*|\boldsymbol{\mu}^*, \boldsymbol{x}^n) \times \quad (3.5) \\ & \quad p(\boldsymbol{\tau}^*|\beta^*, \boldsymbol{\mu}^*, \boldsymbol{x}^n) \times p(\boldsymbol{w}^*|\boldsymbol{\tau}^*, \beta^*, \boldsymbol{\mu}^*, \boldsymbol{x}^n), \quad (3.6) \end{aligned}$$

where $\boldsymbol{\Xi}^*$ could be taken to be the mean of the main Gibbs run. The following is a procedure to estimate the value of the posterior density evaluated at $\boldsymbol{\Xi}^*$:

1. The draws from the full Gibbs run are used to estimate

$$\begin{aligned} p(\boldsymbol{\xi}^*|\boldsymbol{x}^n) &= \int \prod_{j=1}^{k} MVN_d(\xi_j^*|\phi_{j,2}, A_{j,22}) p(\boldsymbol{u}^n, \boldsymbol{z}^n, \boldsymbol{\Sigma}, \beta, \boldsymbol{a}_0|\boldsymbol{x}^n) d\boldsymbol{u}^n d\boldsymbol{z}^n d\boldsymbol{\Sigma} d\beta d\boldsymbol{a}_0 \\ &\sim \frac{1}{G} \sum_{g=1}^{G} \prod_{j=1}^{k} MVN_d(\xi_j^*|\phi_{j,2}^{(g)}, A_{j,22}^{(g)}), \quad (3.7) \end{aligned}$$

where $\phi_{j,2}$ and $A_{j,22}$ are introduced in the previous section.

2. The draws from the reduced Gibbs run with fixed $\boldsymbol{\xi}^*$ are used to estimate

$$\begin{aligned} p(\boldsymbol{a}_0^*|\boldsymbol{\xi}^*, \boldsymbol{x}^n) &= \int \prod_{j=1}^{k} N(a_{j,0}^*|c_j, s_j) I(a_{j,0} < \min_{i \in N_j} u_i) b_j^{-1} \quad (3.8) \\ & \quad p(\boldsymbol{u}^n, \boldsymbol{z}^n, \boldsymbol{\Sigma}, \beta|\boldsymbol{\xi}^*, \boldsymbol{x}^n) d\boldsymbol{u}^n d\boldsymbol{z}^n d\boldsymbol{\Sigma} d\beta \quad (3.9) \end{aligned}$$

where

$$b_j = \Phi(\min_{i \in N_j} u_i), \quad c_j = \phi_{j,1} + A_{j,12} A_{j,22}^{(-1)} (\xi_j^* - \phi_{j,2}), \quad s_j = A_{j,11} - A_{j,12} A_{j,22}^{(-1)} A_{j,21}$$

described in the previous section. Then, we can estimate $p(\boldsymbol{a}_0^*|\boldsymbol{\xi}^*, \boldsymbol{x}^n)$ by

$$\frac{1}{G} \sum_{g=1}^{G} \int \prod_{j=1}^{k} N(a_{j,0}^*|c_j^{(g)}, s_j^{(g)}) / b_j^{(g)},$$

where $b_j^{(g)}$, $c_j^{(g)}$ and $s_j^{(g)}$ are sampled conditional on $\boldsymbol{x}^n$ and $\boldsymbol{\xi}^*$.

3. The draws from another reduced Gibbs run with fixed $\boldsymbol{\mu}^*$ are used to estimate

$$p(\beta^*|\boldsymbol{\mu}^*, \boldsymbol{x}^n) = \int Wish_{d+1}(\beta^*|2g_p + 2kr_p, (2h_p + 2\sum_{j=1}^{k} \tau_j)^{(-1)}) \quad (3.10)$$

$$p(\boldsymbol{u}^n, \boldsymbol{z}^n, \boldsymbol{\Sigma}|\boldsymbol{\mu}^*, \boldsymbol{x}^n) d\boldsymbol{u}^n d\boldsymbol{z}^n d\boldsymbol{\Sigma}, \quad (3.11)$$

which is estimated by

$$\frac{1}{G}\sum_{g=1}^{G} Wish_{d+1}(\beta^*|2g_p + 2kr_p, (2h_p + 2\sum_{j=1}^{k} \tau_j^{(g)})^{(-1)})$$

4. The draws from another subsequent reduced Gibbs run with fixed $\boldsymbol{\mu}^*$ and β^* are used to estimate

$$p(\boldsymbol{\Sigma}^*|\beta^*, \boldsymbol{\mu}^*, \boldsymbol{x}^n) = \int \prod_{j=1}^{k} Wish_{d+1}(\Sigma_j^*|2r_p + n_j, [2(\beta^* + 0.5\sum_{i \in N_j}(y_i - \mu_j^*)(y_i - \mu_j^*)^\top)]^{(-1)})$$

$$p(\boldsymbol{u}^n, \boldsymbol{z}^n|\beta^*, \boldsymbol{\mu}^*, \boldsymbol{x}^n) d\boldsymbol{u}^n d\boldsymbol{z}^n, \quad (3.12)$$

which is estimated by

$$\frac{1}{G}\sum_{g=1}^{G}\prod_{j=1}^{k} Wish_{d+1}(\Sigma_j^*|2r_p + n_j^{(g)}, [2(\beta^* + 0.5\sum_{i \in N_j^{(g)}}(y_i - \mu_j^*)(y_i - \mu_j^*)^\top)]^{(-1)})$$

5. Finally, the draws from the subsequent reduced Gibbs run with fixed $\boldsymbol{\Sigma}^*$, $\boldsymbol{\mu}^*$ and β^* are used to estimate

$$p(\boldsymbol{w}^*|\boldsymbol{\Sigma}^*, \beta^*, \boldsymbol{\mu}^*, \boldsymbol{x}^n) = \int Dirich(\boldsymbol{w}^*|a + n_1, ..., a + n_k)$$

$$p(\boldsymbol{u}^n, \boldsymbol{z}^n|\boldsymbol{\Sigma}^*, \beta^*, \boldsymbol{\mu}^*, \boldsymbol{x}^n) d\boldsymbol{u}^n d\boldsymbol{z}^n, \quad (3.13)$$

which is estimated by

$$\frac{1}{G}\sum_{g=1}^{G} Dirich(\boldsymbol{w}^*|a + n_1^{(g)}, ..., a + n_k^{(g)}).$$

3.5 Application

3.5.1 Iris data

The example data set here is the part of Iris data, where there are 2 classes, Iris setosa, Iris versicolor and 2 variables,

$$X_1 = \text{ sepal length }, \; X_2 = \text{ sepal width}$$

Figure 16 displays sepal length on the horizontal axis and sepal width on the vertical axis. The circles are the class of Iris setosa and the cross indicates Iris virginica.

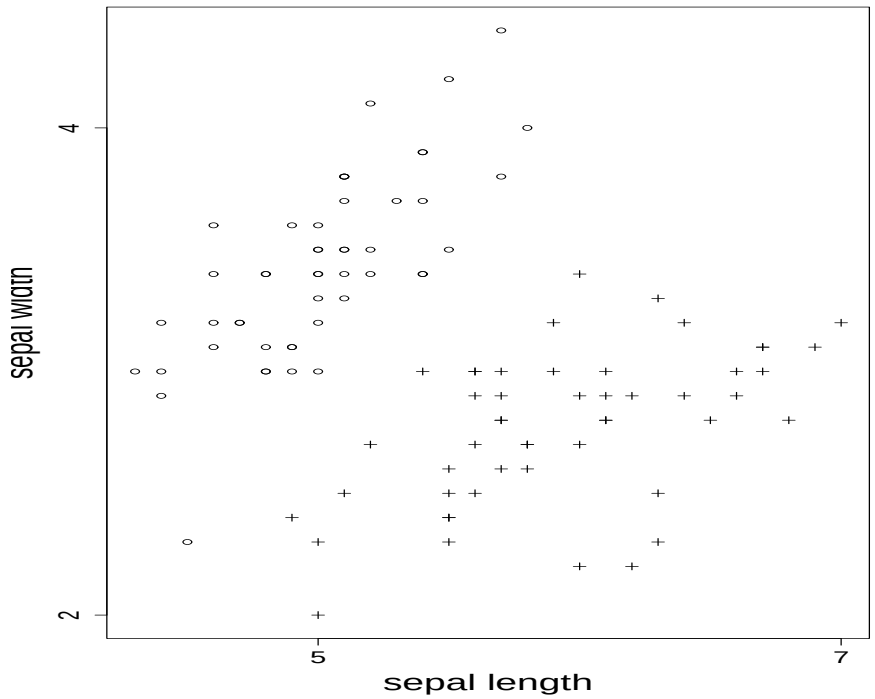

Figure 16: Scatter plot of Iris data. Sepal length on the horizontal axis and sepal width on the vertical axis. The circles are the class of Iris setosa and the cross indicates Iris virginica.

One circle in the bottom left corner is located much closer to the Iris virginica than Iris setosa. This figure shows that each of two groups seems to be approximately

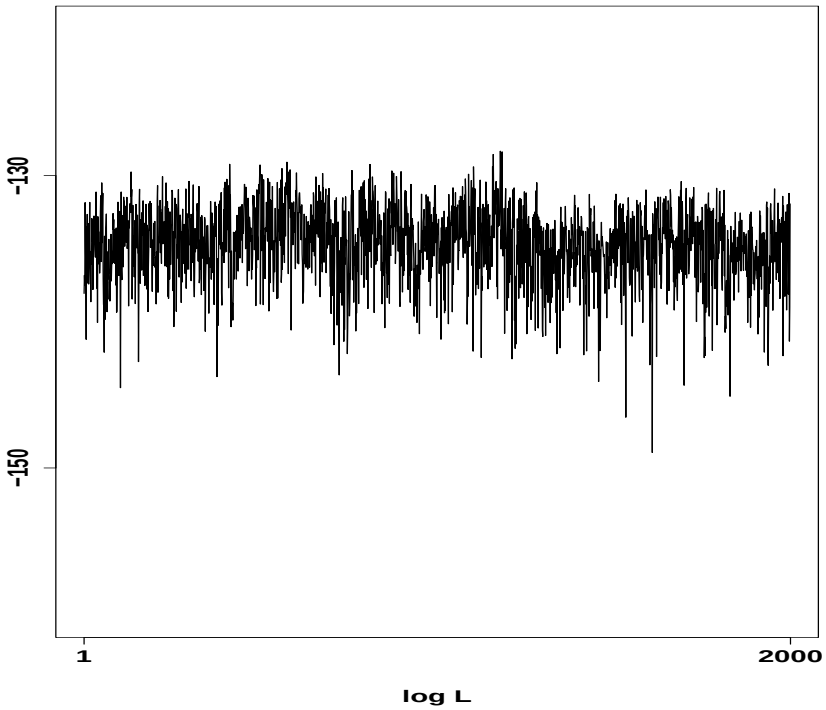

Figure 17: Plot of log likelihood.

normally distributed. The analysis with multivariate skew-normal distribution is done and will be expected to display the values of the skewness parameter close to 0. The analysis under the multivariate normal distribution can be easily done using ghe software MCLUST developed by Fraley and Raftery (1999). From both analyses, the observation corresponding to the outlying circle is wrongly classified and the other 99 observations are correctly grouped.

The simulation results using multivariate skew-normal mixture are displayed in the following figures. The log likelihood per iteration is displayed in the Figure 17.

Figure 18 displays the results for the mixing proportion W_1 and w_2. Each row corresponds to each group, Iris setosa and Iris versicolor with the iteration plot on the first column and its smoothed density estimates on the second column.

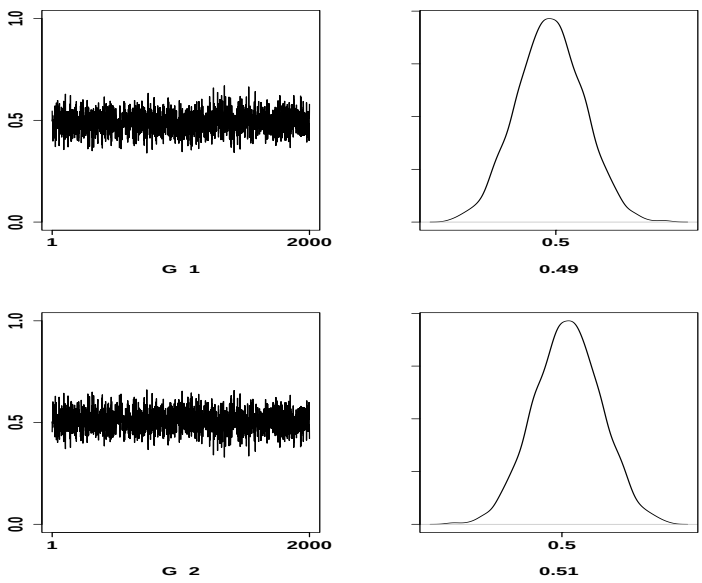

Figure 18: Iteration plot and density estimate of mixing proportions of 2000 samples after 2000 burn-in.

Other parameters including skewness, location and scale parameters are displayed in Figure 19. First two rows are for sepal length and sepal width of Iris setosa and last two rows are for those of Iris versicolor. The numbers under each plot is the mean of 2000 Gibbs samples after 200 burn-in. It is observed that the estimates of skewness parameter are all close to 0 as expected.

The estimates of location parameter for the first group, Iris setosa, is $(5.15, 3.28)$ and those for the Iris versicolor is (5.66,2.97). The estimates from MCLUST which produces (5.02,3.45) and (5.90,2.76), respectively.

The variance estimate from MCLUST are $\begin{pmatrix} 0.12 & 0.09 \\ 0.09 & 0.12 \end{pmatrix}$ for Iris setosa and $\begin{pmatrix} 0.30 & 0.09 \\ 0.09 & 0.10 \end{pmatrix}$ for Iris versicolor. The estimates of scale matrix in the multivariate

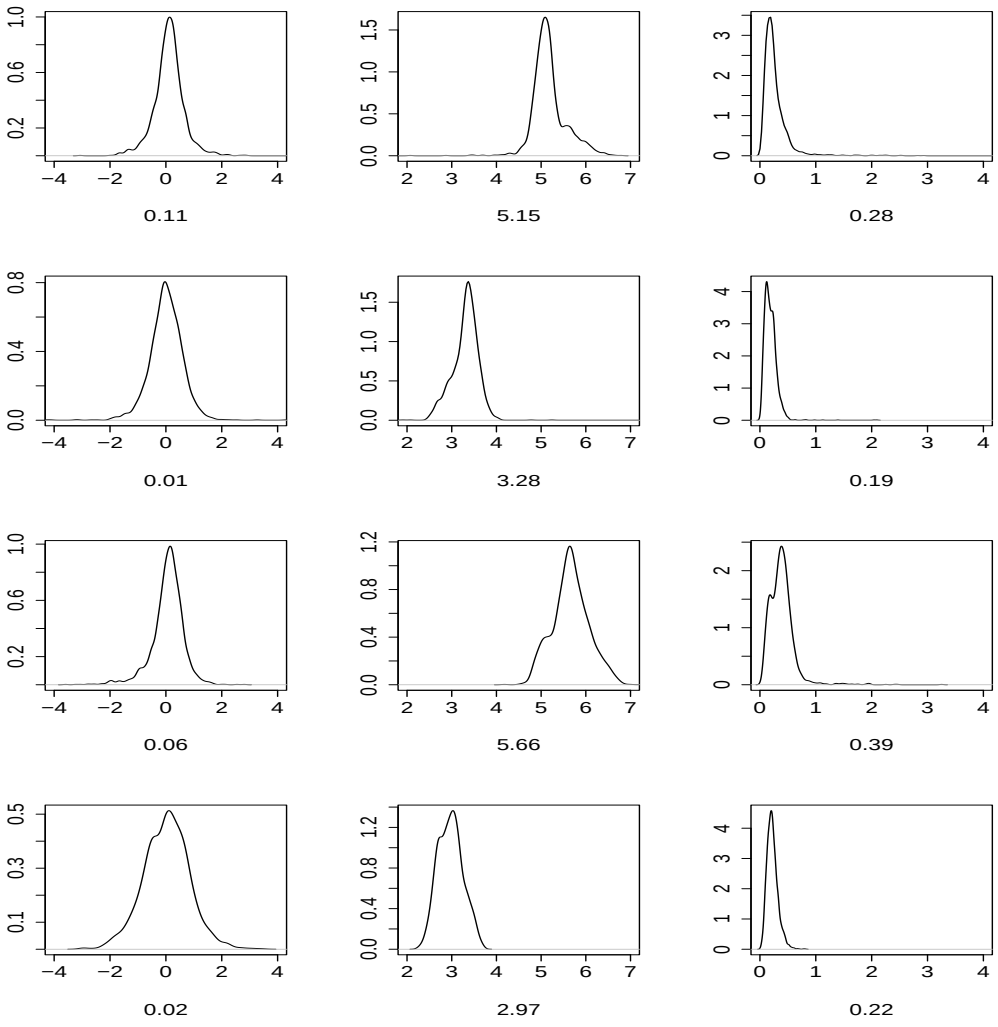

Figure 19: Smoothed density estimate of skewness parameter λ (column 1), location parameter ξ (column 2) and diagonal elements of scale parameter Ω (column 3). First two rows (sepal length and sepal width) are for Iris setosa and last two rows for Iris versicolor. Each number listed under plots is the mean of 2000 samples after 2000 burn-in.

skew-normal distribution are $\begin{pmatrix} 0.28 & 0.00 \\ 0.00 & 0.19 \end{pmatrix}$ for Iris setosa and $\begin{pmatrix} 0.39 & -0.00 \\ -0.00 & 0.22 \end{pmatrix}$ for Iris versicolor.

From the results, we note that the multivariate skew-normal mixture could be applied even to the normally distributed data. The number of parameters fitted are much greater than that of normal mixture. In the next example, we can see the advantage taken by adding the skewness parameter.

3.5.2 Crab data

We consider now the blue crab data set which has been analyzed by Peel and McLachlan (2000). The data set can be easily downloaded from R package MASS. There are 100 observations, 50 for male and 50 for female. Each specimen has measurements on the width of the frontal lip FL, the rear width RW, and length along the midline CL and the maximum width CW of the carapace, and the body depth BD in mm. Figure 20 displays the scatter plot of RW denoted by X_1 and CL by X_2.

Peel and McLachlan (2000) reported 19 of misclassification when a mixture of two normal homoscedastic components is fitted. A fewer miscalssification number of 17 has been reported without restrictions on the scale matrices of multivariate normal mixture model. They continued to fit the mixture model of multivariate t-distribution which is argued to be robust to some outliers resulting in 18 misclassification. The multivariate skew-normal mixture results in 14 misclassification, which yields 20 % decrease in errors.

3.5.3 Simulation

We generated 100 samples from a two-dimensional multivariate skew-normal distribution. The software MCLUST, which uses BIC (Schwarz, 1978), picks up two com-

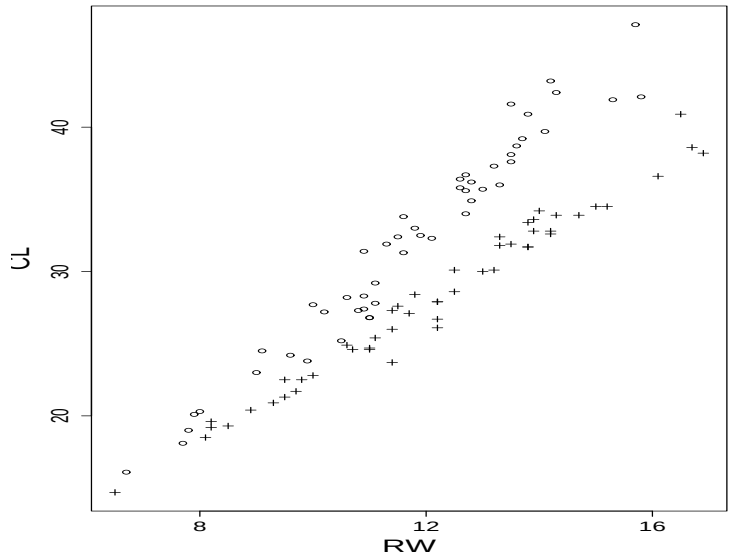

Figure 20: Scatter plot of blue crab data. The rear width RW on the horizontal axis and the midline CL on the vertical axis. The circles are for Male and the cross for Female.

ponents of multivariate normal mixture model. In Figure 21, the scatter plot is displayed.

Instead of BIC we calculate the Bayes factor based on Chib (1995) for choice of number of components among the multivariate skew-normal models. We obtain -375.37 in log scale of the marginal likelihood for one component of skew-normal distribution and -438.53 for mixture of two skew-normal ones, which supports one component rather than 2 components.

3.5.4 Yeast cell data

In this section, we consider the yeast cell data analyzed using multivariate normal mixture model in Yeung et al. (2001a). The yeast cell cycle data (Cho et al., 1998) showed the fluctuation of expression levels of approximately 6000 genes over two cell cycles (17 time points). Yeung et al. (2001a) analyzed the 5-phase criterion

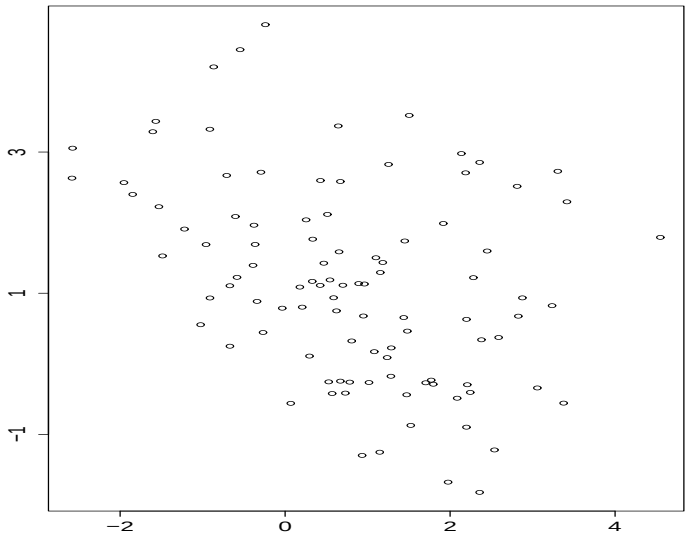

Figure 21: Scatter plot of 100 samples from a 2 dimensional skew-normal distribution.

subset which consists of 384 genes peaking at different time points corresponding to the five phases of cell cycle. Here we expect clustering results to approximate the five phases. To verify multivariate normality they employed the marginal test, the bivariate test, and the radius test (Aitchison, 1986), the skewness and kurtosis test(Mardia, 1970) and the maximum likelihood estimation of the transformation parameters (Yeung et al., 2001b). According to the supplemental web-site (http://faculty.washington.edu/kayee/model), they determined the need for a log transformation. The results of Aitchison tests on the log scale seems satisfactory at the supplemental web-site. But the skewness test on the log scale shows that 3 clusters of genes out of 5 different clusters do not make any improvement in terms of the p-values and the other 2 clusters of genes show little improvement. The estimates of power transformation parameters are made on 5 different classes resulting in different power transformations ranging .14 and .22 for all 5 classes. Instead of taking the different transformations on each of 5 classes or any number between .14

and .22 which would make little sense, they took log transformation on all 5 classes. Even the log transformation does not pass all the tests they considered. Another transformation they considered is the standardization to mean 0 and standard deviation 1. For each gene i, there are 17 observations. The standardization is taken for each gene i, $i = 1, ..., 384$. For each i, there are 17 observations, which are highly correlated. The classes in the data set are based on peak times of the five phases of cell cycle, and so the classes capture the 'general pattern' across the experiments and not the absolute expression levels of the genes. Standardization captures this information better than log transformation. The results of multivariate normality on the standardized data set given at the supplemental web-site displays a more distance to joint normality than log transformation. To incorporate such abnormality, we fit the multivariate skew-normal model to the data in the standardized data set as well as the log transformation conditional on $k = 3, 4, 5$. As a model choice criterion we use Bayes factor described in the previous section. For the purpose of comparison to the results from Yeung et al. (2001a), we also calculate the adjusted Rand indices (Yeung et al., 2001b) , which estimates the measure of agreement even when comparing partitions with different number of clusters.

In the previous chapter, it has been described how to calculate the Bayes factor, which needs to obtain the marginal likelihood as well as the specification of prior model probability. Here the model indication is the number of components in the mixture model. Assuming that the prior probabilities of different models are equally distributed, the Bayes factor is fully decided by the marginal likelihood. Therefore, the implementation of the method developed in Chib (1995), which has been shown in the previous section, estimates the marginal likelihood. First, for the log transformed

data, we have

$$\begin{aligned}\log m(\boldsymbol{x}^n|k=3) &= -5657.055\\ \log m(\boldsymbol{x}^n|k=4) &= -5574.029\\ \log m(\boldsymbol{x}^n|k=5) &= -5751.901,\end{aligned}$$

which supports the model with $k = 4$. Yeung et al. (2001a) reported that BIC analysis selects the model with 5 clusters, which would be expected because of the heavier tail of the skew-normal distribution than that of the normal distribution. For the standardized data, we have the same result, supporting the model with clusters 4 The estimates of the marginal likelihood are

$$\begin{aligned}\log m(\boldsymbol{x}^n|k=3) &= -6905.78\\ \log m(\boldsymbol{x}^n|k=4) &= -6805.84\\ \log m(\boldsymbol{x}^n|k=5) &= -6862.83.\end{aligned}$$

The result by BIC does not agree with the result by the adjusted Rand indices in Yeung et al. (2001a) in the standardized data set. BIC chooses the 5 cluster under the model assuming the same covariance across the clusters but its adjusted Rand index peaks at 7 and take almost same values from 3 to 7 clusters. Changing the assumption on the covariance structure yields the peak at 4 clusters in most of case, which does not agree with the result from the choice by BIC but agrees with the results by skew-normal model. The adjusted Rand indices for the different k mixture of multivariate skew-normal distribution are .4124 for $k = 3$, .4791 for $k = 4$, and .3721 for $k = 5$, which shows the peak at $k = 4$.

3.6 Summary

It is shown that the multivariate skew-normal distribution can be an alternative to the transformation for the analysis of multivariate data with a certain amount of skewness. Its advantage is presented that the analysis and interpretation of the result is easily performed on the original scale of the data. The multivariate skew-normal mixture model is flexible enough to make little difference in the results on the data where multivariate normality is assumed, for example, Iris data. The skewness parameters are estimated closely to zero in the case. When the heavy tailed distribution is assumed as in the crab data example where the multivariate t-distribution has been considered (Peel and McLachlan, 2000), the multivariate skew-normal mixture model shows its usefulness in terms of miss classification rate. Bayes factor conditional on the number of components in the mixture model is proposed for the model determination in the multivariate skew-normal mixture model following the algorithm by Chib(1995). In the example of yeast cell data, the result of the number of components basedn on the marginal likelihood is shown to be not sensitive to the transformation when the multivariate skew-normal distribution is considered.

CHAPTER IV

BAYESIAN MODEL AVERAGING FOR HETEROGENEOUS FRAILTY

4.1 Introduction

Analysis of multivariate failure times entails incorporating the dependence among the observed times into the proportional hazard model. The proportional hazard frailty model, denoted by PHFM hereafter, treats the dependence between multivariate failure times using the unobserved common frailty, which is assumed to follow a specific type of distribution such as gamma and log–normal distribution. To be specific, in the PHFM, the $j-$th recurrent time of the $i-$th subject has the hazards rate at time t as

$$\lambda\big(t \mid v_{ij}, x_{ij}\big) = v_{ij} \cdot \lambda_0\big(t\big) \cdot \exp\big(x_{ij}^T \cdot \beta\big),$$

where $\lambda_0(t)$ is a baseline hazards function, x_{ij}s are covariates and v_{ij} is a vector of random effects (frailty). The frailty v_{ij} is often assumed to be homogeneous which implies the frailty does not vary over time or different covariates: e.g.

$$v_i = v_{i1} = \cdots = v_{im}, \tag{4.1}$$

$$g\big(v_i \big| X_{i1}, X_{i2}, \ldots, X_{im}\big) = g\big(v_i\big), \tag{4.2}$$

where (4.1) and (4.2) implies the homogeneity over time and that over covariates, respectively.

The above two homogeneous assumptions are, however, easily broken in practice. First, when repeatedly measured survival times are observed, it is natural to assume that two adjacent survival times are more dependent of (correlated to) each other

than those two which are apart to each other in time. However, the frailty model assuming (4.1) only allows a positive constant (not varying over time) dependency (correlation) among repeatedly measured survival times. Second, the violation to the assumption (4.2) has long been discussed in the literature of the generalized linear model (GLM) (see the mean variance joint model in McCullough and Nelder (1989)).Although assuming the homogeneity and a specific type of distribution on frailty has been the premise in most works so far, several recent studies showed that different frailty distributions induce quite different dependence structures (Shih and Louis, 1995; Glidden, 1999). The mis-specification in frailty distribution may result in the bias of regression coefficient estimates (Section 10.4.2 Lancaster (1996)). Hence, one should pay careful attention to such aspects especially when the observations exhibit a strong pattern of heterogeneity; for instance, in the non–trivial presence of extraordinary large or small survival times.

To date, much effort has been made to lessen the assumptions on the frailty. First, to lessen the distributional assumption, many previous works have broadened the class of frailty distributions. Especially several heavy tail distributions have been proposed to cope with the unusual frailty estimates. For example, for the kidney data presented in McGilchrist and Aisbett (1991), the frailty estimates from the PHFM with log–normal frailty raise the doubt on the heavy tail frailty distribution. In accordance with these observations, several wider classes of frailty distributions have been proposed, and most of them were on the heavy tail frailty distributions; Sahu and Dey (2004) introduced a log–skewed $t-$ distribution; Ravishanker and Dey (2000) introduced a mixture of positive stable distributions. Second, to overcome the shortcomings from the homogeneity over time, the models with stochastically varying frailty have been proposed by several authors. Some interesting works among many of them are; the dynamic gamma frailty model by Yue and Chan (1996) where the

multiplicative random walk is assumed for varying frailty; the autoregressive (AR) frailty model by Yau and McGilchrist (1998) where the frailty moves according to the AR process; and the similar AR model is considered by Diggle (1988) in the regression model. It should be remarked that the heterogeneous frailty can be easily confused with the heavy tail frailty, in the sense that the heterogeneous frailty, the mixture of homogeneous frailty distribution, has a heavier tail than that of each homogeneous component.

In this chapter, we study the regression model for the variance components in the PHFM through the kidney data example. In the kidney data analysis, it is conjectured by several authors that the individual effects of male group have a larger variance than those of female group (see p640 in Qiou, Ravishanker, and Dey, 1999). Motivated by such observations, we extend the mean variance joint model in the GLM (or the multi–level regression model) to the frailty model and denote it as the multi–level frailty model (MLFM). However, as we will see in the analysis of the kidney data, the observed data often hard to provide any statistical significance between the homogeneous frailty model and the heterogeneous frailty model. For such model uncertainty from frailty distribution, we propose a fully Bayesian approach with the reversible jump Markov chain Monte Calro (MCMC) by Green (1995) to select the model automatically between the PHFM with homogeneous frailty and that with heterogeneous frailty. Thus, the estimate of the regression coefficient ignores the model uncertainty from the frailty distribution in the sense that it averages between the model with homogeneous frailty and that with heterogeneous frailty.

A brief outline of the chpater is as follows. In Section 4.2, we review the analysis of the kidney data in the literature and introduce the model we consider (MLFM). Section 4.3 introduces the Markov chain Monte Carlo procedure to estimate the proposed model. Section 4.4 analyzes the kidney data and implements a simulation

study to see the performance of the proposed procedure for the magnitude of the heterogeneity and the sample size. Finally, Section 4.5 concludes the chapter with discussions of the computing time, possible extensions to more general models, and the extension to the accelerated failure time models.

4.2 Kidney Infection Data and Multi–level Frailty Model

McGilchrist and Aisbett (1991) reported data as to the recurrence times (in days) of infections of 38 kidney patients from insertion of catheter until it has to be removed owing to infection. Three covariates are observed for each patient (sex, age, and disease type), but age and disease type are omitted in the analysis as their effect on the infection time was shown to be insignificant in previous work (McGilchrist and Aisbett, 1991)

To date, many different proportional hazards frailty models have been applied to the kidney data and several authors pointed out the potential heterogeneity in frailty distribution between different gender groups. Among many of them, Qiou et al. (1999) addressed that the frailties for the male are rather irregularly distributed with a larger variance than those of the female. Subsequently, Rabishanker and Dey (2000) proposed the PHFM with a mixture of positive stable distribution as a potential remedy to the above heterogeneous frailty, but they did not use the sex information in explaining the heterogeneity. In this section, we explain such heterogeneous frailty using the regression model between the variance of the individual effects and covariates.

4.2.1 Full description of the model

Let t_{ij} be the j-th recurrent survival time of the i-th subject. Then, given v_i, the hazards function of the model is

$$\lambda\big(t_{ij}\big|v_i; x_{ij}\big) = v_i \cdot \lambda_0\big(t_{ij}\big) \exp\big(x_i\beta\big),$$

where x_{ij} is the covariate "sex", β is the regression coefficient, and $\lambda_0(\cdot)$ is the baseline hazards function.

4.2.1.1 Prior description for frailty

Let $\mathbf{M}_1$ denote the model where the individual effect v_i is assumed to be from the Gamma(α, α), where Gamma(α, β) denotes a gamma distribution with mean α/β and variance α/β^2. $\mathbf{M}_2$ denotes the model where the individual effect v_i is from Gamma(α_1, α_1) if $x_{ij} = x_i = 1$, which means the $i-$th subject is female and it is from Gamma(α_0, α_0) if $x_{ij} = x_i = 0$, which means the $i-$th subject is male. The gamma distribution can be replaced into any other frailty distribution including log–normal distribution and positive stable distribution. However, when the log–normal frailty distribution is used, the interpretation may need a special care due to the identifiability between the regression coefficient and the frailty. It will be discussed in Section 4.3.

Further it is assumed that as a prior distribution, α, α_1, and α_0 are independently and identically distributed (IID) Gamma(κ, κ), where κ is a fixed constant to be estimated or to be guessed. Gelman (2005) discussed the choice of distributions of the hyper-parameters in hierarchical models and provided an example whose final estimates strongly depend on the specification of hyper–parameter distribution. We observe that it does not apply to our case (see Section 4.4)

4.2.1.2 Priors for the regression coefficient

A normal prior is put on the regression coefficient, β, with zero mean and and variance b^2, for which we choose 10^3 as in Sahu et al. (1997).

4.2.1.3 Prior description for the baseline hazards function

The time period is divided into J pre-specified intervals, $I_i = \big(s_{i-1}, s_i\big)$ for $i = 1, 2, \cdots, J$, where $0 = s_0 \leq s_1 \leq s_2 \leq \cdots \leq s_J < \infty$. The baseline hazard function $\lambda_0(t)$ is assumed to be piecewise constant as

$$\lambda_0(t) = \lambda_k \qquad \text{if} \quad s_k \leq t \leq s_{k+1}.$$

In this chapter, we assume the piecewise independent gamma distribution as a prior for the baseline hazards function, which assumes $\underline{\lambda} = (\lambda_1, \cdots, \lambda_J)$ is from

$$f\big(\underline{\lambda}\big) = \prod_{k=1}^{J} f\big(\lambda_k\big),$$

where $f(\lambda_k)$ is Gamma(τ_k, τ_k). It is interesting to see that the proposed prior is equivalent to the independent increment gamma process prior in Clayton (1991), when τ_k is proportional to the length of the interval I_k. Although, we do not adapt it in this chapter, there has been considerable amount of efforts on the correlated prior process including Arjas and Gasbarra (1994), and Aslanidou et al. (1998) .

Finally, in this chapter, we empirically determine J as in Qiou et al. (1999), but a random choice of J can be considered using the reversible jump MCMC by Green (1995).

4.2.1.4 Prior distribution between models $\mathbf{M}_1$ *and* $\mathbf{M}_2$

As noted in Section 2.1.1, the model $\mathbf{M}_1$ assumes the homogeneous variance of the frailty distribution and $\mathbf{M}_2$ assumes the heterogeneous variance structure depending

on the covariate ("sex" in this chapter). Prior probability of each model is set to $P(\mathbf{M}_1) = P(\mathbf{M}_2) = 0.5$.

4.2.2 Connection of MLFM to existing models

The considered MLFM can be interpreted as extensions of some existing models (or problems) to survival data.

First, similar models are addressed in the context of the multi–level regression model in the previous literature (Section 5.2 in Heagerty and Zeger, (2000)), but most of them are limited to non–survival data problems. A few of them on survival data are Yau (2001), Maples et al. (2002), and Zhang and Steele (2004). In particular, the random coefficient PHM by Maples et al. (2002), is a special case of our MLFM. Second, in the PHFM, α is the scale parameter of marginal distribution of the survival time as well as that of the frailty distribution. More specifically, in the PHFM with gamma frailty distribution, the marginal distribution of observed survival time is

$$f(t;\beta,\alpha) = \int f\big(t|v;\beta\big) \cdot g(v;\alpha)dv,$$

where $f\big(t|v;\beta\big) = \lambda_0(t)\exp(x^T\beta + \log v)\exp\{-\Lambda_0(t)\exp(x^T\beta + \log v)\}$ and $\Lambda_0(t) = \int_0^t \lambda_0(s)ds$. For example, when $g(v;\alpha)$ is Gamma(α,α) with $\lambda_0(t) = 1$ and $\beta = 0$, the marginal distribution $f(t)$ becomes $\exp\big((\alpha+1)\log\big(\alpha/(\alpha+t)\big)\big)$. Thus, as a model for survival times $\big\{t_{ij}, i = 1,2,\cdots,38,\ j = 1,2\big\}$, it can be considered as an extension of the (mean variance) joint-model of the GLM (see McCullough and Nelder (1989)) to the proportional hazard model. Finally, it also can be considered as an extension of the Behrens–Fisher problem to survival models. The Behrens–Fisher problem concerns the inference on the difference between the means of two normal populations whose ratio of variances is unknown. In our model for the kidney data, we only consider the sex variable in both the model for hazards rate, $\lambda(t_{ij})$, and the

variance components, $\alpha_k,\ k = 0, 1$. Thus, testing the regression coefficient β (of sex) is equivalent to testing the mean of survival times between the male group and female group. Furthermore, the ratio of the variances between survival times of male group and those for female group is unknown.

4.3 Full Conditional Distributions and MCMC

Recall that $\mathbf{M}_1$ and $\mathbf{M}_2$ denote the PHFM with homogeneous frailty and that with heterogeneous frailty, respectively. Then, the overall sampler has three components: (1) the sampler in $\mathbf{M}_1$, (2) the sampler in $\mathbf{M}_2$, and (3) reversible jump MCMC between the space of $\mathbf{M}_1$ and $\mathbf{M}_2$.

4.3.1 Sampling algorithms in $\mathbf{M}_1$

Given v_i, the $j-$th recurrent time of the $i-$th subject has a constant hazard of

$$h_{ij} = \lambda_k \cdot \eta_{ij} \cdot v_i$$

in the $k-$th interval ($k = 1, 2, ..., J$) with $\eta_{ij} = \exp(x'_{ij}\beta)$. If the subject has survived beyond the $k-$th interval, i.e., $t_{ij} > s_k$ for s_k defined in Section 2.1.2, the likelihood contribution is

$$\exp\big(-\lambda_k \cdot \Delta_k \cdot \eta_{ij} \cdot v_i\big),$$

where $\Delta_k = s_k - s_{k-1}$. If the subject has failed or is censored in the $k-$th interval, i.e., $s_{k-1} < t_{ij} \leq s_k$, then the likelihood contribution is

$$(\lambda_k \cdot \eta_{ij} \cdot v_i)^{\delta_{ij}} \cdot \exp\big(-\lambda_k \cdot (t_{ij} - s_{k-1}) \cdot \eta_{ij} \cdot v_i\big),$$

where $\delta_{ij} = 1$ if t_{ij} is not censored; otherwise, it is 0.

Let $D = (X, \delta, t, v)$ denote the complete data and $D_{\text{Obs}} = (X, \delta, t)$ denote the

observed data. Then, the complete data likelihood becomes

$$
\begin{aligned}
&l\big(\beta, \underline{\lambda}, \underline{v}, \alpha \big| D_{\text{Obs}}\big) \\
&\quad \propto \prod_{i=1}^{n} \prod_{j=1}^{m_i} \Bigg\{ \Bigg(\prod_{k=1}^{g_{ij}} \exp\big(-\lambda_k \cdot \Delta_k \cdot \eta_{ij} \cdot v_i \big) \Bigg) \Big(\lambda_{g_{ij}+1} \cdot \eta_{ij} \cdot v_i \Big)^{\delta_{ij}} \\
&\qquad\qquad \times \exp\Big(-\lambda_{g_{ij}+1} \cdot \big(t_{ij} - s_{g_{ij}} \big) \cdot \eta_{ij} \cdot v_i \Big) \Bigg\},
\end{aligned}
$$

where g_{ij} is such that $t_{ij} \in (s_{g_{ij}}, s_{g_{ij}+1}] = I_{g_{ij}+1}$ and $\underline{v} = \big(v_1, \ldots, v_n\big)$. Now we specify the full conditional distributions of unknowns for the MCMC implementation.

First, with the prior Gamma(α, α), the full conditional distribution of v_i for each $i = 1, 2, ..., n$, becomes

$$
\mathrm{P}\big(v_i \big| \beta, \underline{\lambda}, \alpha, D_{\text{Obs}}\big) \sim \text{Gamma}\big(\alpha + \sum_{j=1}^{m_i} \delta_{ij}, \alpha + S_i\big),
$$

where

$$
\begin{aligned}
S_i &= \sum_{j=1}^{m_i} \eta_{ij} \cdot \Big(\sum_{k=1}^{g_{ij}} \lambda_k \cdot \Delta_k + \lambda_{g_{ij}+1} \cdot \big(t_{ij} - s_{g_{ij}} \big) \Big) \\
&= \sum_{j=1}^{m_i} e^{x'_{ij} \cdot \beta} \Big(\sum_{k=1}^{g_{ij}} \lambda_k \cdot \Delta_k + \lambda_{g_{ij}+1} \cdot \big(t_{ij} - s_{g_{ij}} \big) \Big).
\end{aligned}
$$

Second, with the prior $\pi\big(\alpha\big) = \text{Gamma}(\kappa, \kappa)$, the full conditional of α is

$$
\mathrm{P}\big(\alpha \big| \beta, \underline{\lambda}, \underline{v}, D_{\text{Obs}}\big) \propto \frac{\eta^{n\alpha}}{\Gamma(\alpha)^n} \exp\Big\{ -\eta \cdot \sum_i \big(v_i - \log v_i \big) \Big\} \times \pi\big(\alpha\big).
$$

The Metropolis–Hastings algorithm is implemented to get a sample from the posterior distribution.

Third, with the prior $\pi\big(\beta\big)$, the full conditional of β is

$$
\mathrm{P}\big(\beta \big| \underline{\lambda}, \underline{v}, \alpha, D_{\text{Obs}}\big) \propto \exp\Big(\sum_{i=1}^{n} \sum_{j=1}^{m_i} \delta_{ij} \cdot x'_{ij} \cdot \beta \Big) \cdot \exp\Big(-\sum_{i=1}^{n} v_i \cdot S_i \Big) \times \pi\big(\beta\big).
$$

The sample from the posterior distribution can be obtained using the Metropolis–Hastings algorithm again.

Fourth, with the prior of piecewise independent gamma distribution, say $\pi\big(\lambda_k\big) = \text{Gamma}(\tau_k, \tau_k)$, the full conditional of λ_k is

$$\text{P}\big(\lambda_k\big|\beta, \underline{v}, \alpha, D_{\text{Obs}}\big) \sim \text{Gamma}\big(d_k + \tau_k, V_k + \tau_k\big),$$

where d_k is the number of failure times occurred in the interval $I_k = (s_{k-1}, s_k]$ and

$$V_k = \sum_{(i,j)\in\mathcal{R}_k} \Delta_k \eta_{ij} v_i + \sum_{(i,j)\in\mathcal{D}_k} (t_{ij} - s_{k-1}) \cdot \eta_{ij} \cdot v_i. \tag{4.3}$$

R_k and D_k in (4.3) is the set of indexes of the subjects who survive longer than s_{k-1} and those of subjects who died in the interval I_k, respectively.

4.3.2 Sampling algorithms in $\mathbf{M_2}$

The only difference between the sampler of $\mathbf{M}_2$ from that of $\mathbf{M}_1$ is the prior distribution of v, which is, in $\mathbf{M}_2$,

$$\pi\big(v\big|Z = k\big) = \text{Gamma}\big(\alpha_k, \alpha_k\big) \qquad \text{for} \qquad k = 1 \text{ or } 0,$$

where Z indicates the sex of a subject. Thus, the full conditional distributions of β and λ_k are same as those in $\mathbf{M}_1$.

Let $N_k,\ k = 1, 0$ be the index set of subjects whose Z value is k and $n_k = |N_k|$, the number of subjects in N_k. Then, the full conditional distribution of α_k is

$$\begin{aligned}\text{P}\big(\alpha_k\big|\beta, \underline{\lambda}, \underline{v}, D_{\text{Obs}}\big) &\propto \prod_{i\in N_k} \pi\big(v_i\big|\alpha_k\big)\pi\big(\alpha_k\big) \\ &= \frac{\alpha_k^{n_k\alpha_k}}{\Gamma(\alpha_k)^{n_k}} \cdot \exp\Big\{ -\alpha_k \sum_{i\in N_k} (v_i - \log v_i)\Big\} \times \pi\big(\alpha_k\big).\end{aligned}$$

The samples from the full conditionals are obtained using the Metropolis-Hastings algorithm as in $\mathbf{M}_1$.

4.3.3 Reversible MCMC between $\mathbf{M}_1$ *and* $\mathbf{M}_2$

Among existing model choice techniques, we consider to use the reversible jump MCMC algorithm. It enables us to simultaneously estimate the posterior probabilities of several models under consideration and the parameters conditional on a specific model.

Green (1995) developed the reversible jump algorithm, which generalizes the Metropolis-Hastings algorithm into the dimension varying situation. Consider the finite mixture model, for example. The dimension of the parameter space depends on the number of components. When the number of components in the model is also unknown, the ordinary MCMC algorithm cannot be directly implemented since the dimension of all the unknowns is not fixed. The reversible jump MCMC algorithm is designed in such a way that the sampler moves across different dimensions. The following shows how to implement the reversible jump MCMC into the models we have considered.

There are 4 types of possible moves between $\mathbf{M}_1$ and $\mathbf{M}_2$; the move within $\mathbf{M}_1$; the move within $\mathbf{M}_2$; the move from $\mathbf{M}_1$ to $\mathbf{M}_2$; and the move from $\mathbf{M}_2$ to $\mathbf{M}_1$. The moves within each $\mathbf{M}_1$ and $\mathbf{M}_2$ are discussed in the previous section and the moves from one to the other model will be discussed here. Before describing the details of each move, it should be remarked that, unlike mixture cases (Richardson and Green, 1997), the RJMCMC in our problem does not need to consider the allocation of the subjects to each group during the move from $\mathbf{M}_1$ to $\mathbf{M}_2$ because the allocation is predetermined by the covariate "sex".

First, consider the move from $\mathbf{M}_1$ to $\mathbf{M}_2$. The dimension of the parameter space in $\mathbf{M}_2$ is greater than $\mathbf{M}_1$ by 1 because we have (α_1, α_0) in $\mathbf{M}_2$ for the variance of the frailty depending on the sex but only α in $\mathbf{M}_1$. For dimension matching, we

need an additional continuous random variable. Let w be a random number from the exponential distribution with a rate of $\log 2$ (hence, the median is 1). The candidate values α_1 and α_0, then, can be defined as

$$\alpha_1 = w \cdot \alpha \qquad \text{and} \qquad \alpha_0 = \alpha/w.$$

When $y = \alpha$ and $y' = (\alpha_1, \alpha_0)$, the sampler moves from $\mathbf{M}_1$ to $\mathbf{M}_2$ with probability $\rho = \min(1, A)$, where

$$A = \frac{\pi\big(y'\big|\beta, \underline{\lambda}, \underline{v}, D_{\text{Obs}}\big) r_m\big(y'\big)}{\pi\big(y\big|\beta, \underline{\lambda}, \underline{v}, D_{\text{Obs}}\big) r_m\big(y\big) q\big(w\big)} \cdot \left|\frac{\partial y'}{\partial(y, w)}\right|, \tag{4.4}$$

following the equation (7) in Section 3.1 in Richardson and Green(1997). Here, the last term is the Jacobian arising at the transformation $(y, w) \to y'$, that is,

$$\left|\frac{\partial y'}{\partial(y, w)}\right| \;=\; \begin{vmatrix} w & \alpha \\ 1/w & -\alpha/w^2 \end{vmatrix} = \frac{2 \cdot \alpha}{w}$$

and the ratio of the other part for A is

$$\frac{\pi\big(y'\big|\beta, \underline{\lambda}, \underline{v}, D_{\text{Obs}}\big) r_m\big(y'\big)}{\pi\big(y\big|\beta, \underline{\lambda}, \underline{v}, D_{\text{Obs}}\big) r_m\big(y\big) q\big(w\big)} = \frac{\prod_{k=1}^{2} \prod_{i \in N_k} \pi\big(v_i\big|\alpha_k\big) \pi\big(\alpha_k\big)}{\prod_{i=1}^{n} \pi\big(v_i\big|\alpha\big) \pi\big(\alpha\big) q\big(w\big)},$$

where $q(w)$ is the density function of the exponential distribution with the rate of 2 evaluated at w, and $r_m(y)$ is the probability of choosing move type m when in state y, which is 1 in our case.

Next, let us consider the move from $\mathbf{M}_2$ to $\mathbf{M}_1$. When the chain moves from $\mathbf{M}_2$ to $\mathbf{M}_1$, the parameters α and w are directly computed from α_1 and α_0 as:

$$\alpha = \sqrt{\alpha_1 \cdot \alpha_0} \qquad \text{and} \qquad w = \sqrt{\alpha/\alpha_0}.$$

Subsequently, the sampler moves from $\mathbf{M}_2$ to $\mathbf{M}_1$ with probability

$$\min\big(1,\ \mathbf{A}^{-1}\big),$$

with appropriate substitutions in (4.4) .

The posterior probability $p(\mathbf{M}_k|$ Data$)$ is estimated by the relative frequency of the number of iterations the sampler visits $\mathbf{M}_k$.

4.3.4 Estimation of the hyper-parameters

In our full specification, there are three different hyper-parameters b, τ, and κ; b is from the prior distribution of β; τ is from the prior distribution of the baseline hazards; κ is from the prior distribution of the variance of frailty distribution. It is often observed that the posterior distribution is quite sensitive to the specification of the hyper-parameters. In such case, it is more sensible to estimate those parameters in the empirical Bayesian point of view.

Let $Y = \big(T, \delta\big)$ be the observed survival times and the censoring information, and let W be other random components including β, $\underline{\lambda}$, $\underline{v}$ which are not observed. Then, from the conventional missing data arguments,

$$\begin{aligned}
\frac{\partial f\big(Y\big|b,\tau,\kappa\big)}{\partial(b,\tau,\kappa)} &= \frac{\partial \int f\big(Y,W\big|b,\tau,\kappa\big)dW}{\partial(b,\tau,\kappa)} = \int \frac{\partial f(Y,W\big|b,\tau,\kappa\big)}{\partial(b,\tau,\kappa)}dW \\
&= \int \frac{\partial \log f\big(Y,W\big|b,\tau,\kappa\big)}{\partial(b,\tau,\kappa)} f\big(Y,W\big|b,\tau,\kappa\big)dW \\
&= f\big(Y\big|b,\tau,\kappa\big)\int \frac{\partial \log f\big(Y,W\big|b,\tau,\kappa\big)}{\partial(b,\tau,\kappa)} f\big(W\big|Y,b,\tau,\kappa\big)dW.
\end{aligned}$$

Thus, maximizing marginal log–likelihood function $\log f\big(Y\big|\kappa\big)$ is equivalent to maximizing the expected complete log–likelihood function (expectation is with respect to the posterior distribution)

$$\int \log f\big(Y,W\big|b,\tau,\kappa\big)\cdot f\big(W\big|Y,b,\tau,\kappa\big)\,dW. \tag{4.5}$$

Finally, for each κ, (4.5) can be approximated using the posterior samples $\big\{W^{(i)}\big\}_{i=1}^{B}$ by

$$\int \log f\big(Y,W\big|b,\tau,\kappa\big) f\big(W\big|Y,b,\tau,\kappa\big)dW \approx \frac{1}{B}\sum_{i=1}^{B} \log f\big(Y,W^{(i)}\big|b,\tau,\kappa\big).$$

In the kidney data example from the next section, the posterior distribution of $\mathbf{M}_2$ and others are not much sensitive to the specification of hyper-parameters when

τ is sufficiently small ($\tau \leq 0.1$) (see Figure 22). Finally, we set $b^2 = 1000$, $\kappa = 0.05$, and $\tau = 0.1$.

4.4 Examples

4.4.1 Kidney data analysis

In this section, we apply the proposed MLFM to the kidney infection data in McGilchrist and Aisbett (1991). $20,000$ samples were generated from the posterior distribution. The fast convergence of the sampler can be checked from the log-likelihood values of each Gibbs sample. Finally, $10,000$ samples are selected after $10,000$ burn-in period for the inference of β, $\underline{\lambda}$, $\underline{v}$, and α. Hereinafter, the estimate refers the posterior sample mean.

The frailty v_is (for $i = 1, 2, \cdots, 38$) are assumed to be from Gamma$\big(\alpha, \alpha\big)$ in $\mathbf{M}_1$ and they are assumed to be from Gamma$\big(\alpha_1, \alpha_1\big)$ or Gamma$\big(\alpha_0, \alpha_0\big)$ relying on the sex in $\mathbf{M}_2$. Each of the above αs are assumed to be from Gamma$\big(\kappa, \kappa\big)$. Before we report our results, it should be pointed out that in both the positive stable frailty model and the gamma frailty model, the frailty estimates of male are more variable than those of female (see Figure 23 and p. 640 in Qiou et al. (1999)).

In our analysis, the frailty estimates in each group approximately have a mean of one and, in the heterogeneous frailty model, the variability of the frailty estimates differs slightly between male and female groups, but it is not much apparent as in Qiou et al. (1999). (see Figure 24). The frailty estimate of each patient is presented in Table 6 for $\mathbf{M}_1$ and $\mathbf{M}_2$ separately.

The posterior probability of the heterogeneous model $\mathbf{M}_2$ was estimated by the proportion of the iterations, where the sampler stayed in $\mathbf{M}_2$. It was computed as 10.06%. Thus, the observed data did not provide any statistically significant preference between the homogeneous frailty model and the heterogeneous frailty model.

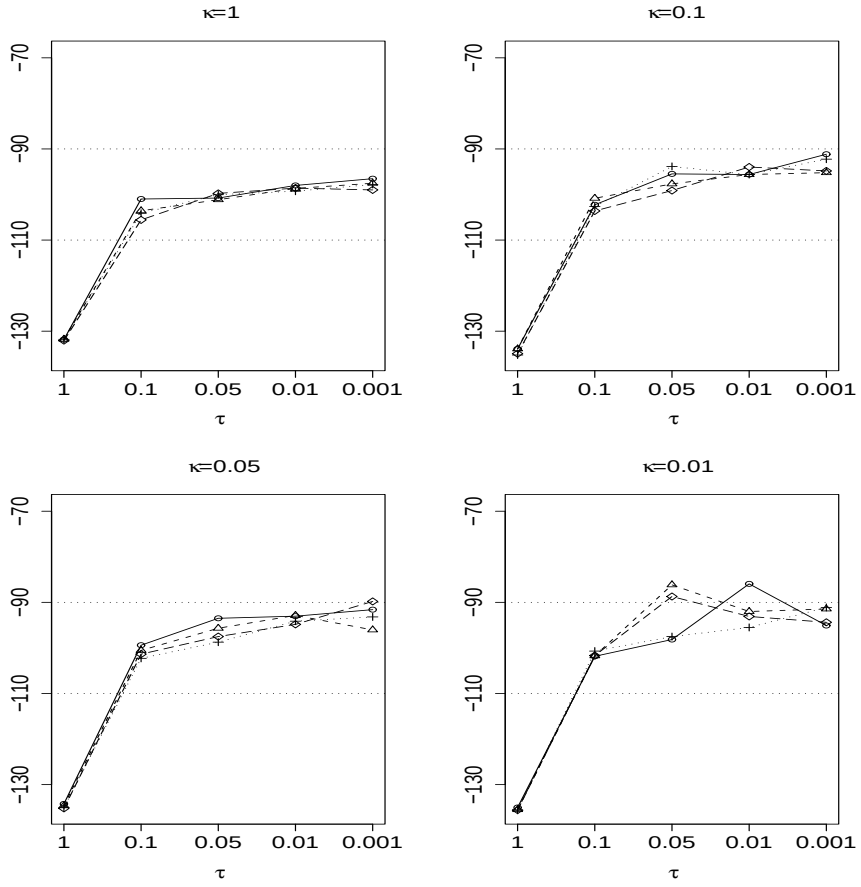

Figure 22: Plots of mean of penalized log likelihood from MCMC samples for different values of hyper-parameters: (a) $\kappa = 1$ (b) $\kappa = 0.1$ (c) $\kappa = 0.05$ (d) $\kappa = 0.01$. In each plot, the circle indicates the case for $b^2 = 10$, triangle for $b^2 = 25$, cross for $b^2 = 100$ and diamond for $b^2 = 1000$. The x-axis indicates the inverse of variance of $\pi(\lambda)$, which appears in $G(\tau, \tau)$.

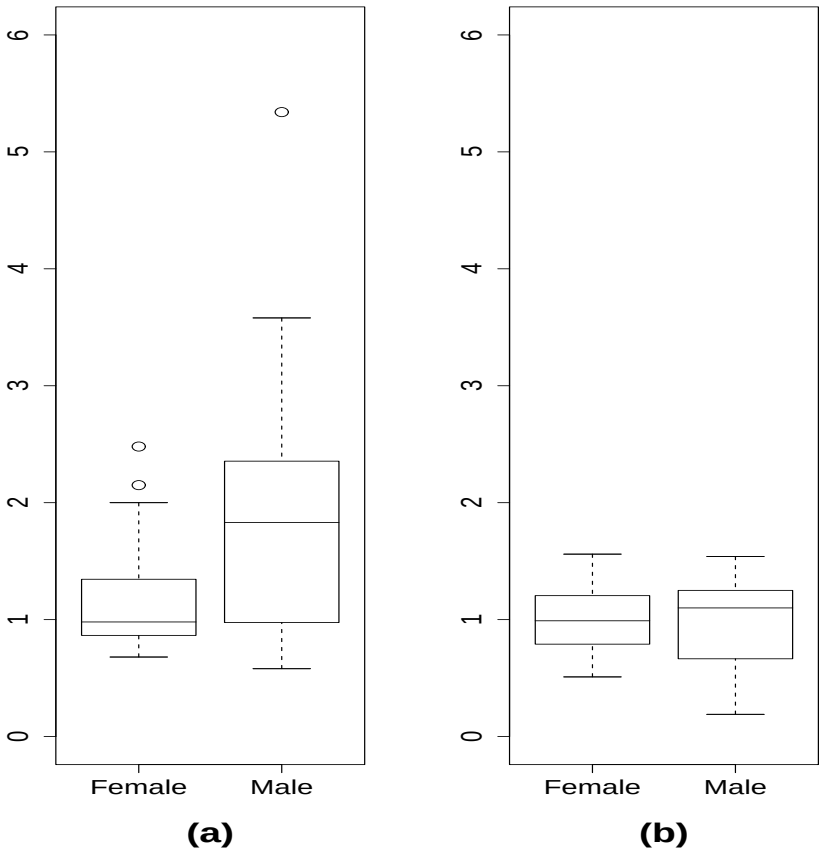

Figure 23: Box plots of posterior means of frailties by sex from two different frailty distributions by Qiou et al. (1999): (a) posterior means assuming positive stable distribution for frailties, (b) posterior means assuming gamma distribution for frailties.

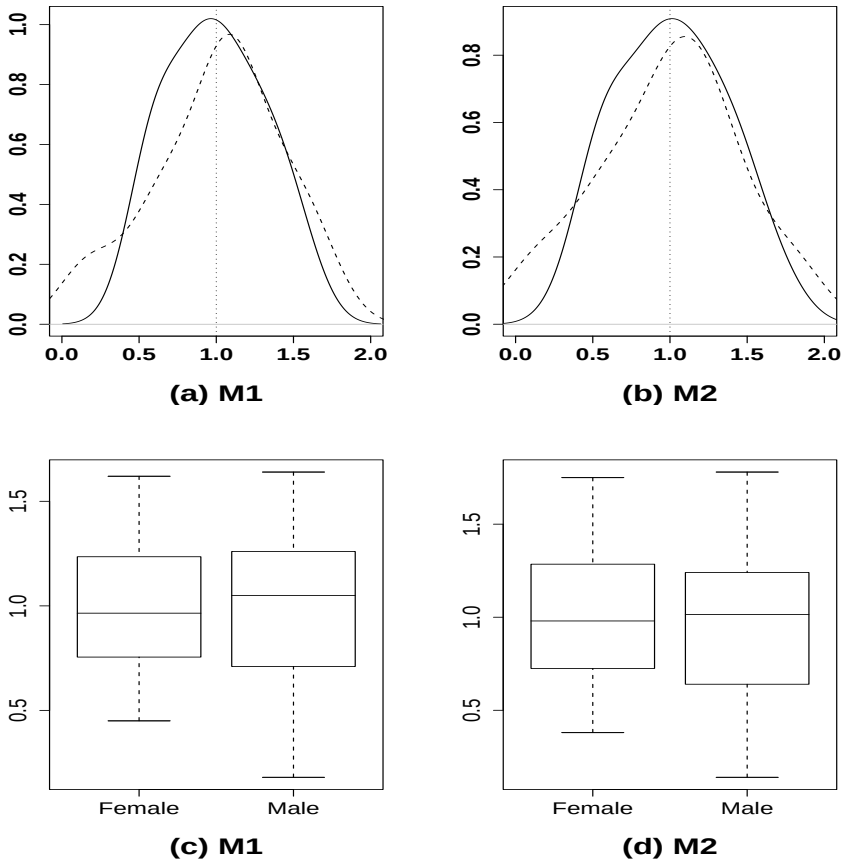

Figure 24: Density plots (first row) and box-plots (second row) of estimated frailties over sex and models. (a) and (b) : Density plot of posterior means of frailties with solid line for female and dashed line for male. (c) and (d) : Box-plots of posterior means of frailties of female and male for each model.

Table 6: Posterior means and standard deviations of frailties by sex for two different models. $\mathbf{M}_1$ denotes the case of homogeneous variance on the frailty, whereas $\mathbf{M}_2$ denotes the case of heterogeneous variance.

M1				M2			
Female		Male		Female		Male	
Mean	SD	Mean	SD	Mean	SD	Mean	SD
1.32	0.78	1.47	0.77	1.31	0.76	1.42	0.79
0.54	0.31	1.02	0.51	0.5	0.31	0.96	0.51
0.95	0.46	1.08	0.54	0.96	0.51	1.07	0.58
0.58	0.34	1.64	0.91	0.54	0.34	1.78	1.08
0.98	0.48	0.58	0.35	1	0.55	0.52	0.35
1.62	0.89	1.15	0.65	1.75	0.96	1.22	0.7
0.95	0.52	0.18	0.18	0.96	0.57	0.14	0.15
1.27	0.65	0.96	0.48	1.37	0.78	0.93	0.53
0.59	0.4	1.26	0.64	0.57	0.43	1.24	0.67
0.45	0.31	0.71	0.42	0.38	0.28	0.64	0.4
0.83	0.41			0.82	0.47		
0.8	0.4			0.78	0.44		
0.63	0.41			0.6	0.44		
1.04	0.58			1.07	0.66		
0.64	0.38			0.61	0.41		
1.49	0.79			1.54	0.98		
1.05	0.58			1.1	0.68		
0.71	0.39			0.67	0.38		
0.95	0.46			0.96	0.51		
1.47	0.78			1.56	0.87		
1.2	0.59			1.26	0.67		
1.42	0.73			1.45	0.79		
1.2	0.69			1.19	0.71		
1.12	0.55			1.17	0.59		
0.85	0.46			0.88	0.53		
1.33	0.67			1.41	0.78		
0.8	0.52			0.82	0.65		
1.13	0.64			1.18	0.77		

Table 7: Posterior means of the regression coefficient of sex and the variances (the inverse of αs) of gamma frailty over two different models: Gamma denotes the estimates from gamma frailty model listed in Table 1 in Qiou et al. (1999). $\mathbf{M}_1$ denotes the case of homogeneous variance on the frailty, whereas $\mathbf{M}_2$ denotes the case of heterogeneous variance.

	Gamma		M1		M2	
Parameter	Posterior mean	Posterior SD	Posterior mean	Posterior SD	Posterior mean	Posterior SD
β	-1.62	.4186	-1.4871	0.5027	-1.6122	0.5452
$1/\alpha$	.3268	.1737	0.4737	0.2961	–	–
$1/\alpha_1$	–	–	–	–	0.5482	0.2869
$1/\alpha_0$	–	–	–	–	0.5702	0.321

In the heterogeneous frailty model ($\mathbf{M}_2$), the variance estimates of the frailty distribution in the male group and the female group were 0.5702 and 0.5482, respectively. However, the variance estimate of the frailty distribution in the homogeneous frailty model ($\mathbf{M}_1$) was estimated as 0.3268. It should be noted that, unexpectedly, this variance estimate was smaller than both 0.5482 and 0.5702 in $\mathbf{M}_2$. This strange phenomena was an outcome of the prior effect to the posterior distribution when the number of the observations were small; it was expected that each of the above variance estimates was the average of the prior information (forced that the variance was one) and the data information (indicated that the variance was smaller than 1). Subsequently, when estimating the variance in $\mathbf{M}_2$, the variance estimates were more influenced by the prior due to the small number of observations. The variance estimates are reported in Table 7.

The baseline hazard function is modelled by a piecewise gamma distribution. The break points of the time axis are chosen as $\{0, 10, 20, 30, 40, 50, 100, 200, 300, 400\}$ as in Qiou et al. (1999) and the prior of the baseline hazards in each interval is chosen independently and identically as Gamma(τ, τ) with $\tau = 0.1$. The estimates are

Table 8: Posterior means of baseline hazard rates denoted by $\lambda_j,\ j = 1, ..., 10$ for the kidney infection data: Gamma denotes the estimates from gamma frailty model listed in Table 1 in Qiou et al. (1999). $\mathbf{M}_1$ denotes the case of homogeneous variance on the frailty, whereas $\mathbf{M}_2$ denotes the case of heterogeneous variance.

	Gamma		$\mathbf{M}_1$		$\mathbf{M}_2$	
Parameter	Posterior mean	Posterior SD	Posterior mean	Posterior SD	Posterior mean	Posterior SD
λ_1	0.0015	0.003	0.0212	0.0112	0.0232	0.0124
λ_2	0.0012	0.0025	0.0377	0.0215	0.0422	0.0255
λ_3	0.0011	0.002	0.0833	0.0464	0.0901	0.0511
λ_4	0.001	0.002	0.0583	0.0439	0.065	0.048
λ_5	0.0012	0.0023	0.015	0.0193	0.0163	0.0209
λ_6	0.0011	0.0024	0.0241	0.0193	0.0277	0.0212
λ_7	0.0011	0.0024	0.0514	0.0409	0.0606	0.0455
λ_8	0.0014	0.0028	0.0426	0.0526	0.0497	0.0545
λ_9	0.0056	0.0051	0.0251	0.0499	0.0307	0.0529
λ_{10}	0.3667	0.1495	0.1594	0.229	0.1982	0.2584

reported in Table 8 and it is also interesting to see the the estimate for the interval $[400, \infty)$ is much larger than those for other intervals as in Qiou et al. (1999).

Finally, β is negatively estimated as -1.56 whose absolute value is larger than that from the homogeneous frailty (-1.4871) and smaller than that from the heterogeneous frailty (-1.6122). the female patients have a lower risk for infection. The posterior distribution of β in $\mathbf{M}_2$ is not much different from that in M_1. The posterior samples of β are plotted in Figure 25. Also, parameter estimates are presented in Table 7.

4.4.2 Simulated data examples

The kidney data analysis in the previous section shows the model uncertainty between the heterogeneous frailty model and its homogeneous counterpart. In this section, we implement a simulation study to investigate the performance of the proposed Bayesian procedure for various magnitudes of the heterogeneity in variance components and

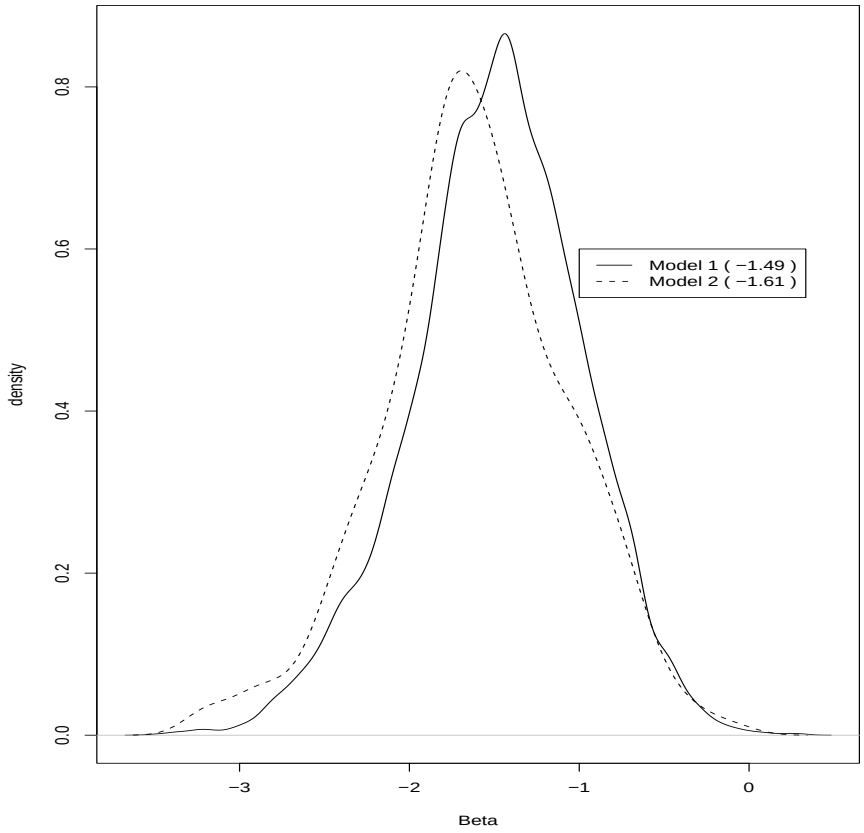

Figure 25: Estimated densities for the regression coefficient β from 20 000 iterations after 10 000 burn-in from two different models. The straight line denotes the estimated density from $\mathbf{M}_1$ and the dashed line from $\mathbf{M}_2$. The values in the parentheses are the estimated posterior means.

Table 9: Means and standard deviations of β estimates from 50 simulations when $n = 50$ and $n = 100$.

		$(\alpha 1, \alpha 2)$			
n	Model	(0.1 , 10)	(0.2 , 5)	(0.5 , 2)	(1 , 1)
50	$\mathbf{M}_1$	2.239 (1.037)	2.463 (0.793)	2.441 (0.713)	2.305 (0.576)
	$\mathbf{M}_2$	2.001 (1.020)	2.279 (0.787)	2.385 (0.679)	2.316 (0.556)
	Model Avg.	2.181 (1.038)	2.427 (0.796)	2.434 (0.710)	2.305 (0.575)
100	$\mathbf{M}_1$	2.477 (0.923)	2.754 (0.512)	2.55 (0.573)	2.327 (0.314)
	$\mathbf{M}_2$	2.255 (0.929)	2.592 (0.510)	2.486 (0.542)	2.330 (0.304)
	Model Avg.	2.389 (0.946)	2.717 (0.518)	2.542 (0.571)	2.328 (0.313)

sample sizes.

The data sets are simulated from the PHFM with gamma frailty distribution. The baseline hazard function is set to be constant over time as $\lambda_0(t) = 0.01$ and the regression coefficient for the single covariate x is $\beta = 2.0$. The number of subjects (sample size) is $n = 50$ or $n = 100$, where the frailty for each subject is from the Gamma(α_1, α_1) or Gamma(α_0, α_0) according to x. We consider 4 choices of (α_1, α_0) having different magnitudes of heterogeneity; $\big(\alpha_1, \alpha_0\big) = \big(0.1, 10\big)$, $\big(0.2, 5\big)$, $\big(0.5, 2\big)$, or $\big(1, 1\big)$.

In the analysis, we use the same prior distributions with those in the kidney example. The baseline hazards function follows a piecewise exponential distribution with rate λ_k, where each λ_k is from Gamma(τ, τ). The regression coefficient β is from the Gaussian distribution with mean 0 and variance b^2. The frailty v_is are assumed to be Gamma$\big(\alpha, \alpha\big)$ in $\mathbf{M}_1$ and Gamma$\big(\alpha_1, \alpha_1\big)$ or Gamma$\big(\alpha_0, \alpha_0\big)$ according to the sex in $\mathbf{M}_2$, where α, α_1, and α_0 follow from Gamma(κ, κ). The hyper-parameters are $b^2 = 1000$, $\kappa = 0.05$, and $\tau = 0.11$ as in the kidney data analysis.

Table 9 reported the average posterior mean of β over 50 data sets. It showed that the model averaging estimate is adaptive in the sense that it is close to the estimate from $\mathbf{M}_1$ (or $\mathbf{M}_2$) when the data is generated from $\mathbf{M}_1$ (or $\mathbf{M}_2$).

Figure 26 plotted the posterior probabilities of $\mathbf{M}_1$ for several different values of variance components and the sample sizes. It showed that the posterior probability of $\mathbf{M}_2$ increased as the frailty became more heterogenerous. However, its probability is still lower than than that of $\mathbf{M}_1$ even the frailty was very heterogeneous (for example, $\big(\alpha_1, \alpha_0\big) = \big(0.1, 10\big)$) with moderate size $n = 100$.

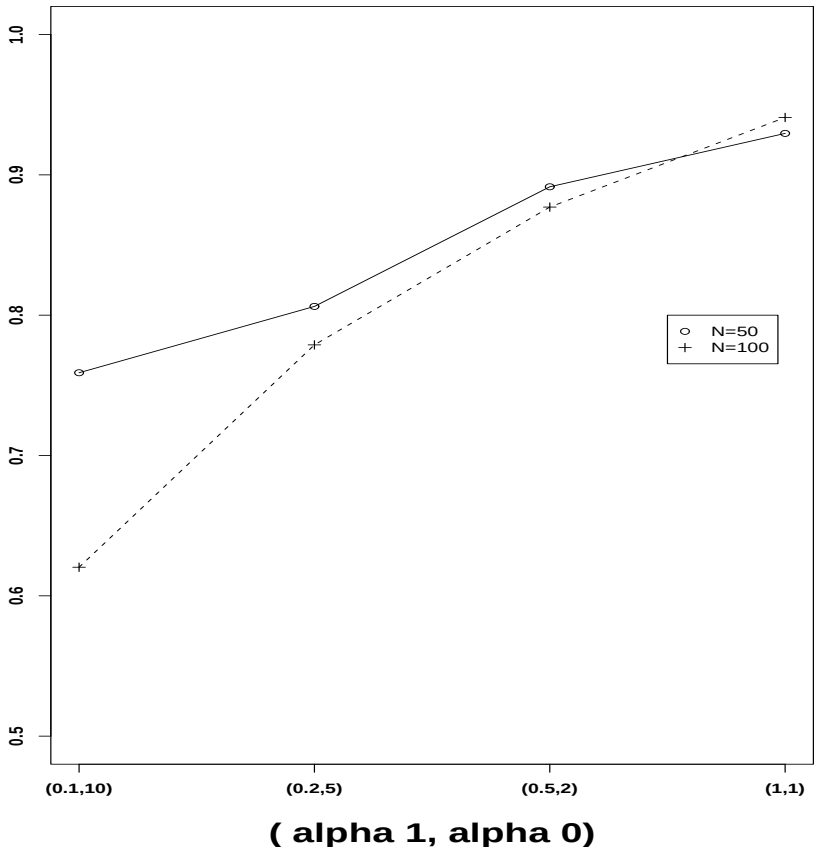

Figure 26: Mean of estimated $P(\mathbf{M}_1|x)$ from 50 simulated data sets.

4.5 Discussion

In this chapter, we consider a regression model for variance components in the PHFM and propose a fully Bayesian procedure with the RJMCMC for the difficulty of model

uncertainty from the frailty distribution. In this section, we conclude the chapter with a few discussions not covered in the main body of the chapter. First, it should be pointed out that the computing time is not much different from the MCMC procedure for the PHFM with homogeneous frailty since the assignment of subject to two groups (subjects has homogeneous frailty in each group) is guided to the covariates (sex in the Kidney data analysis). Second, the procedure of this chapter can straightforwardly be extended to multi–sample problems (the regression covariate is discrete but has more than two different levels) with a more complicate RJMCMC. However, the extension to more general settings such as the model with continuous covariate or the multiple regression model are still to be pursued. Finally, the same issue in other survival models such as the accelerated failure time model with the frailty does not much differ from that in the PHFM.

CHAPTER V

CONCLUSIONS

We have proposed a mixture of skew-normal distribution as an alternative of changing the original scale including the Box-Cox transformation when there exists a certain amount of skewness in the data. The number of components is estimated by the reversible jump MCMC algorithm which reports the posterior probability of the number of components. When there are negative values in the observations, the Box-Cox transformation needs another parameter to shift the data to make them all positive. The amount of shift has been reported to affect the entire estimation procedure and the inference on the number of components.

The analysis on the original scale by using a skew-normal mixture model has been presented in both univariate and multivariate distributions under the existence of skewness. It did not change the scale of the data and did not have any limitation on the range of the data, which produces much easier interpretation of the results. We analyzed the enzyme data with good results in terms of posterior probability of the number of components.

The reversible jump algorithm is applied to proportional hazard model with frailty. The probability of different models considerd is presented for a model determination. From both simulated data and kidney infection data, the sucessful convergence of the algorithm is presented along with Bayesian model averaging procedure to incorporate model uncertainty.

REFERENCES

Aitchison, J. (1986). *The statistical analysis of compositional data*. London: Chapman and Hall.

Arjas, E. and Gasbarra, D. (1994). "Nonparametric Bayesian inference for right–censored survival data." *Statistica Sinica*, 4, 505–524.

Arnold, B. C., Beaver, R. J., Groeneveld, R. A., Meeker, W. Q., and Brooks, S. P. (1993). "The nontruncated marginal of a truncated bivariate normal distribution." *Psychometrika*, 58, 471–478.

Aslanidou, H., Dey, D., and Sinha, D. (1998). "Bayesian analysis of multivariate survival data using Monte Carlo methods." *Canadian Journal of Statistics*, 26, 33–48.

Azzalini, A. (1985). "A class of distributions which includes the normal ones." *Scandinavian Journal of Statistics*, 12, 171–178.

Azzalini, A. and Capitanio, A. (1999). "Statistical applications of the multivariate skew normal distribution." *Journal of the Royal Statistical Society*, Ser. B, 61, 579–602.

Azzalini, A. and Dalla Valle, A. (1996). "The multivariate skew normal distribution." *Biometrika*, 83, 715–726.

Bechtel, Y. C., Bonáiti-Pelliè, C., Poisson, N., Magnette, J., and Bechtel, P. R. (1993). "A population and family study of N-acetyltransferase using caffeine urinary metabolities." *Clininal Pharmaceutical Therapeutics*, 54, 134–141.

Box, G. E. P. and Cox, D. R. (1964). "An analysis of transformation." *Journal of the Royal Statistical Society*, Ser. B, 26, 211–252.

Brooks, S. P., Giudici, P., and Roberts, G. O. (2003). "Efficient construction of reversible jump MCMC proposal distributions (with discussion)." *Journal of the Royal Statistical Society*, Ser. B, 65, 3–57.

Carlin, B. and Chib, S. (1995). "Bayesian model choice via Markov chain Monte Carlo methods." *Journal of the Royal Statistical Society*, Ser. B, 57, 473–484.

Carroll, R. J., Ruppert, D., Crainiceanu, C. M., Tosteson, T. D., and Karagas, M. R. (2004). "Nonlinear and nonparametric regression and instrumental variables." *Journal of the American Statistical Association*, 99, 736–750.

Chen, M., Dey, D. K., and Shao, Q. (1999). "A new skewed link model for dichotomous quantal response data." *Journal of the American Statistical Association*, 94, 1172–1186.

Chib, S. (1995). "Marginal likelihood from Gibs output." *Journal of the American Statistical Association*, 90, 1313–1321.

Chib, S. and Jeliazkov, I. (2001). "Marginal likelihood from the Metropolis-Hastings output." *Journal of the American Statistical Association*, 90, 1313–1321.

Cho, R. J., Campbell, M. J., Winzeler, E. A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T. G., Gabrielian, A., Landsman, E., Lockhart, D. J., and Davis, R. W. (1998). "A genome-wide transcriptional analysis of the mitotic cell cycle." *Molecular Cell*, 2, 65–73.

Copas, J. B. and Li, H. G. (1997). "Inference for non-random samples." *Journal of the Royal Statistical Society*, Ser. B, 59, 55–95.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). "Maximum likelihood from incomplete data via the EM algorithm (with discussion)." *Journal of the Royal Statistical Society*, Ser. B, 39, 1–38.

Denison, D. G. T., Holmes, C. C., Mallick, B. K., and Smith, A. F. M. (2002). *Bayesian methods for nonlinear classfication and regression*. Chichester: Wiley.

Diebolt, J. and Robert, C. P. (1994). "Estimation of finite mixture distributions through Bayesian sampling." *Journal of the Royal Statistical Society*, Ser. B, 56, 363–375.

Diggle, P. J. (1988). "An approach to the analysis of repeated measurements." *Biometrics*, 44, 959–971.

Fraley, C. and Raftery, A. E. (1999). "MCLUST: Software for model-based cluster analysis." *Journal of Classification*, 16, 297–306.

— (2002). "Model-based clustering, discriminant analysis, and density estimation." *Journal of the American Statistical Association*, 97, 611–631.

Gelman, A. (2005). "Prior distributions for variance parameters in hierarchical models." *Bayesian Analysis*, (to appear).

Genton, M. (2004). *Skew-elliptical distributions and their applications: A journey beyond normality*. Boca Raton, FL: Chapman & Hall/CRC.

Genton, M. and Loperfido, N. (2005). "Generalized skew-elliptical distributions and their quadratic forms." *Annals of the Institute of Statistical Mathematics*, (to appear).

Glidden, D. A. (1999). "Checking the adequacy of the gamma frailty model for multivariate failure times." *Biometrika*, 86, 381–393.

Green, P. J. (1995). "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination." *Biometrika*, 82, 711–732.

Green, P. J., Hjort, N. L., and Richardson, S. (2003). *Highly structured stochastic systems*. Oxford: Oxford University Press.

Gutierrez, R. G., Carroll, R. J., Wang, N., Lee, G.-H., and Taylor, B. H. (1995). "Analysis of tomato root initiation using a normal mixture distribution." *Biometrics*, 51, 1461–1468.

Heagerty, P. J. and Zeger, S. L. (2000). "Marginalized multilevel models and likelihood inference." *Statistical Science*, 15, 1–26.

Jennison, C. (1997). "Discussion on Bayesian analysis of mixtures with an unknown number of components (by Richardson and Green)." *Journal of the Royal Statistical Society*, Ser. B, 59, 778–779.

Kass, R. and Raftery, A. E. (1995). "Bayes factors." *Journal of the American Statistical Association*, 90, 773–395.

Kim, H.-M. and Mallick, B. K. (2004). "A Bayesian prediction using the skew Gaussian distribution." *Journal of Statistical Planning and Inference*, 120, 85–101.

Lancaster, T. (1996). *The econometric analysis of transition data*. New York: Cambridge University Press.

Liseo, B. and Loperfido, N. (2005). "A note on reference priors for the scalar skew-normal distribution." *Journal of Statstical Planning and Inference*, (to appear).

Liu, J. S., Shang, J. L., Palumbo, M. J., and Lawrence, C. E. (2003). "Bayesian clustering with variable and transformation selections." In *Bayesian statistics 7*,

eds. J. M, Bernardo, M, Bayarri, J. O, Berger, and A. P, Dawid, 249–275. Oxford: Oxford University Press.

MacLean, C. J., Morton, N. E., Elston, R. C., and Yee, S. (1976). "Skewness in commingled distributions." *Biometrics*, 32, 695–699.

Manly, B. F. J. (1976). "Exponential data transformations." *Statistician*, 25, 37–42.

Maples, J. J., Murphy, S. A., and Axinn, W. G. (2002). "Two level proportional hazards models." *Biometrics*, 58, 754–763.

Mardia, K. V. (1970). "Measures of multivariate skewness and kurtosis with applications." *Biometika*, 57, 519–530.

McCullough, P. and Nelder, J. A. (1989). *Generalized linear models*. New York: Chapman and Hall.

McGilchrist, C. A. and Aisbett, C. W. (1991). "Regression with frailty in survival analysis." *Biometrics*, 47, 461–466.

McLachlan, G. and Peel, D. (2000). *Finite mixture models*. New York: Wiley.

McLachlan, G. J., Bean, R. W., and Peel, D. (2002). "A mixture model-based approach to the clustering of microarray expression data." *Bioinformatics*, 18, 413–422.

Nobile, A. (2004). "On the posterior distribution of the number of components in a finite mixture." *Annals of Statistics*, 32, 2044–2073.

O'Hagan, A. and Leonard, T. (1976). "Bayes estimation subject to uncertainty about parameter constraints." *Biometrika*, 63, 201–202.

Pearson, K. (1894). "Contributions to the theory of mathematical evolution." *Philosophical Transcations of the Royal Society of London A*, 185, 71–110.

— (1895). "Contributions to the theory of mathematical evolution, II:skew variaton." *Philosophical Transcations of the Royal Society of London A*, 186, 343–414.

Peel, D. and McLachlan, G. J. (2000). "Robust mixture modelling using the t distribution." *Statistics and Computing*, 10, 339–348.

Qiou, Z., Ravishanker, N., and Dey, D. (1999). "Multivariate survival analysis with positive stable frailties." *Biometrics*, 55, 637–644.

Ravishanker, N. and Dey, D. (2000). "Modeling multivariate survival models with a mixture of positive stable frailties." *Methodology and Computing in Applied Probability*, 2, 293–308.

Richardson, S. and Green, P. J. (1997). "On Bayesian analysis of mixtures with an unknown number of components (with discussion)." *J. R. Statist. Soc. B*, 59, 731–792.

Roberts, C. (1966). "A correlation model useful in the study of twins." *J. Am. Statist. Ass.*, 61, 1184–1190.

Sahu, S. K. and Dey, D. K. (2004). "On multivariate survival models with a skewed frailty and a correlated baseline hazard process." In *Skew-elliptical distributions and their applications: A journey beyond normality*, ed. M. G, Genton, 321–338. Boca Raton, FL: CRC/Chapman & Hall.

Sahu, S. K., Dey, D. K., Aslanidou, H., and Sinha, D. (1997). "A Weibull regression model with gamma frailty for multivariate survival data." *Lifetime Data Analysis*, 3, 123–137.

Schork, N. J. and Schork, M. A. (1988). "Skewness and mixtures of normal distributions." *Communications in Statistics-Theory and Methods*, 17, 3951–3969.

Schork, N. J., Weder, A. B., and Schork, M. A. (1990). "On the asymmetry of biological frequency distributions." *Genetic Epidemiology*, 7, 427–446.

Schwarz, G. (1978). "Estimating the dimension of a model." *Annals of Statistics*, 6, 461–464.

Shih, J. H. and Louis, T. A. (1995). "Inferences on the association parameter in copula models for bivariate survival data." *Biometrics*, 51, 1384–1399.

Stephens, M. (1997). *Bayesian methods for mixtures of normal distributions*. Univ. Oxford: Ph.D. dissertation.

— (2000). "Bayesian analysis of mixture models with an unknown number of components-an alternative to reversible jump methods." *Annals of Statistics*, 28, 40–74.

Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985). *Statistical analysis of finite mixture distributions*. New York: Wiley.

Tukey, J. (1957). "On the comparative anatomy of transformations." *Ann. Math. Statist.*, 28, 602–632.

Yau, K. K. W. (2001). "Multilevel models for survival analysis with random effects." *Biometrics*, 57, 96–102.

Yau, K. K. W. and McGilchrist, C. A. (1998). "ML and REML estimation in survival analysis with time dependent correlated frailty." *Statistics in Medicine*, 17, 1201–1213.

Yeung, K. Y., Fraley, C., Murua, A., Raftery, A. E., and Ruzzo, W. L. (2001a). "Model-based clustering and data transformations for gene expression data." *Bioinformatics*, 17, 977–987.

— (2001b). "Model-based clustering and data transformations for gene expression data." *Technical Report UW-CSE-01-04-02*, Department of Computer Science and Engineering, University of Washington.

Yue, H. and Chan, K. S. (1996). "A dynamic frailty model for multivariate survival data." *Biometrics*, 53, 785–793.

Zhang, W. and Steel, F. (2004). "A semiparametric multilevel survival model." *Journal of the Royal Statistical Society C*, 53, 387–404.

VITA

Ilsung Chang was born in Incheon, Korea in July, 12th, 1970. He is the son of Mankyu Chang and Youngshin Ghil and is also a brother of Minsun and Jeong Geum. He received a B.S. degree in mathematics in 1994 and an M.S. degree in statistics under the supervision of Dr. Shinsup Cho in 1996 both from Seoul National University, Seoul, Korea. In 1999 he came to Texas A&M University to pursue a Ph.D. in statistics. His recent research interests include mixture model, model-based clustering, Bayesian computation, survival analysis and statistical consulting. His permanent address is as follows:

Ilsung Chang
Yonsoo Goo, Oklyun Dong, Hyundai APT 406-601,
Incheon, Korea, 401-060

or, he can be reached at ilsung.chang@gmail.com.